THE INTERNET ENCYCLOPEDIA

Volume 1
A–F

Hossein Bidgoli
Editor-in-Chief
California State University
Bakersfield, California

WILEY
John Wiley & Sons, Inc.

THE INTERNET ENCYCLOPEDIA

Volume 1
A–F

Hossein Bidgoli
Editor-in-Chief
California State University
Bakersfield, California

WILEY
John Wiley & Sons, Inc.

This book is printed on acid-free paper. ♾

Copyright © 2004 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, e-mail: permcoordinator@wiley.com.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. The publisher is not engaged in rendering professional services, and you should consult a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services please contact our Customer Care Department within the U.S. at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books. For more information about Wiley products, visit our web site at www.Wiley.com.

Library of Congress Cataloging-in-Publication Data:

The Internet encyclopedia / edited by Hossein Bidgoli.
p. cm.
Includes bibliographical references and index.
ISBN 0-471-22202-X (CLOTH VOL 1 : alk. paper) – ISBN 0-471-22204-6 (CLOTH VOL 2 : alk. paper) – ISBN 0-471-22203-8 (CLOTH VOL 3 : alk. paper) – ISBN 0-471-22201-1 (CLOTH SET : alk. paper)
1. Internet–Encyclopedias. I. Bidgoli, Hossein.
TK5105.875.I57I5466 2003
004.67′8′03–dc21
2002155552

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

To so many fine memories of my brother, Mohsen, for his
uncompromising belief in the power of education.

About the Editor-in-Chief

Hossein Bidgoli, Ph.D., is Professor of Management Information Systems at California State University. Dr. Bidgoli helped set up the first PC lab in the United States. He is the author of 43 textbooks, 27 manuals, and over four dozen technical articles and papers on various aspects of computer applications, e-commerce, and information systems, which have been published and presented throughout the world. Dr. Bidgoli also serves as the editor-in-chief of *Encyclopedia of Information Systems*.

Dr. Bidgoli was selected as the California State University, Bakersfield's 2001–2002 Professor of the Year.

Editorial Board

Eric T. Bradlow
The Wharton School of the University of Pennsylvania

Kai Cheng
Leeds Metropolitan University, United Kingdom

Mary J. Cronin
Boston College

James E. Goldman
Purdue University

Marilyn Greenstein
Arizona State University West

Varun Grover
University of South Carolina

Ephraim R. McLean
Georgia State University, Atlanta

David E. Monarchi
University of Colorado, Boulder

Raymond R. Panko
University of Hawaii at Manoa

Norman M. Sadeh
Carnegie Mellon University

Judith C. Simon
The University of Memphis

Vasja Vehovar
University of Ljubljana, Slovenia

Russell S. Winer
New York University

Contents

Volume 1

Volume 2

Volume 3

Chapter List by Subject Area

Applications

Design, Implementation, and Management

Electronic Commerce

Foundation

Wireless Internet and E-commerce

Contributors

Tarek Abdelzaher
University of Virginia
Web Quality of Service

Charles Abzug
James Madison University
Linux Operating System

Patricia Adams
Education Resources
Strategic Alliances

Carol A. Akerelrea
Colorado State University
Usability Testing: An Evaluation Process for Internet Communications

Gary C. Anders
Arizona State University West
Online Auctions

Amy W. Apon
University of Arkansas
Public Networks

Pierre A. Balthazard
Arizona State University West
Groupware

Ashok Deo Bardhan
University of California, Berkeley
Real Estate

Joey Bargsten
University of Oregon
Multimedia

Hossein Bidgoli
California State University, Bakersfield
Computer Literacy
Internet Literacy

Gerald Bluhm
Tyco Fire & Security
Patent Law

Robert J. Boncella
Washburn University
Secure Sockets Layer (SSL)

J. Efrim Boritz
University of Waterloo, Canada
XBRL (Extensible Business Reporting Language): Business Reporting with XML

Sviatoslav Braynov
State University of New York at Buffalo
Data Mining in E-commerce
Personalization and Customization Technologies

Randy M. Brooks
Millikin University
Online Publishing

Colleen Brown
Purdue University
History of the Internet

Tara Brown-L'Bahy
Harvard University
Distance Learning (Virtual Learning)

Linda S. Bruenjes
Lasell College
Internet2

Gerard J. Burke
University of Florida
Supply Chain Management

L. Jean Camp
Harvard University
Peer-to-Peer Systems

Charles J. Campbell
The University of Memphis
Java

Janice E. Carrillo
University of Florida
Inventory Management

Michael A. Carrillo
Oracle Corporation
Inventory Management

Lillian N. Cassel
Villanova University
Wireless Application Protocol (WAP)

J. Cecil
New Mexico State University
Virtual Enterprises

Haluk Cetin
Murray State University
Geographic Information Systems (GIS) and the Internet

Henry Chan
The Hong Kong Polytechnic University, China
Consumer-Oriented Electronic Commerce

C. Janie Chang
San José State University
Public Accounting Firms

Camille Chin
West Virginia University
Cybercrime and Cyberfraud

T. Matthew Ciolek
The Australian National University, Australia
Online Religion

Timothy W. Cole
University of Illinois at Urbana-Champaign
Visual Basic Scripting Edition (VBScript)

Fred Condo
California State University, Chico
Cascading Style Sheets (CSS)

David E. Cook
University of Derby, United Kingdom
Standards and Protocols in Data Communications

Marco Cremonini
Università di Milano, Italy
Disaster Recovery Planning

Mary J. Cronin
Boston College
Mobile Commerce

Jaime J. Dávila
Hampshire College
Digital Divide

Chris Dede
Harvard University
Distance Learning (Virtual Learning)

Victoria S. Dennis
Minnesota State Bar Association
Law Firms

Lynn A. DeNoia
Rensselaer Polytechnic Institute
Wide Area and Metropolitan Area Networks

Nikhilesh Dholakia
University of Rhode Island
Gender and Internet Usage
Global Diffusion of the Internet

Ruby Roy Dholakia
University of Rhode Island
Gender and Internet Usage
Global Diffusion of the Internet

Vesna Dolnicar
University of Ljubljana, Slovenia
Benchmarking Internet

Rich Dorfman
WebFeats! and Waukesha County Technical College
Extensible Markup Language (XML)

Magda El Zarki
University of California—Irvine
Wireless Internet

Larry P. English
Information Impact International, Inc.
Information Quality in Internet and E-business Environments

Roman Erenshteyn
Goldey-Beacom College
ActiveX

Ray Everett-Church
ePrivacy Group, Inc.
Privacy Law
Trademark Law

Patrick J. Fahy
Athabasca University
Web-Based Training

Gerald R. Ferrera
Bentley College
Copyright Law

Daniel R. Fesenmaier
University of Illinois at Urbana–Champaign
Travel and Tourism

C. Patrick Fleenor
Seattle University
Feasibility of Global E-business Projects

Marcia H. Flicker
Fordham University
Securities Trading on the Internet

Immanuel Freedman
Dr. Immanuel Freedman, Inc.
Video Compression

Borko Furht
Florida Atlantic University
Interactive Multimedia on the Web

Jayne Gackenbach
Athabasca University, Canada
Health Issues

Alan Gaitenby
University of Massachusetts, Amherst
Online Dispute Resolution

Bruce Garrison
University of Miami
Online News Services (Online Journalism)

G. David Garson
North Carolina State University
E-government

Roger Gate
IBM United Kingdom Ltd., United Kingdom
Electronic Funds Transfer

Mario Giannini
Code Fighter, Inc., and Columbia University
C/C++

Julia Alpert Gladstone
Bryant College
International Cyberlaw

Mary C. Gilly
University of California, Irvine
Consumer Behavior

Robert H. Goffman
Concordia University
Electronic Procurement

James E. Goldman
Purdue University
Firewalls

Sven Graupner
Hewlett-Packard Laboratories
Web Services

Robert H. Greenfield
Computer Consulting
Circuit, Message, and Packet Switching

Ulrike Gretzel
University of Illinois at Urbana–Champaign
Travel and Tourism

Paul Gronke
Reed College
Politics

Jim Grubbs
University of Illinois at Springfield
E-mail and Instant Messaging

Mohsen Guizani
Western Michigan University
Wireless Communications Applications

Jon Gunderson
University of Illinois at Urbana–Champaign
Universally Accessible Web Resources: Designing for People with Disabilities

Babita Gupta
California State University, Monterey Bay
Global Issues

Louisa Ha
Bowling Green State University
Webcasting

Kirk Hallahan
Colorado State University
Online Public Relations

Diane M. Hamilton
Rowan University
Business-to-Consumer (B2C) Internet Business Models

Robert W. Heath Jr.
The University of Texas at Austin
Digital Communication

Geert Heijenk
University of Twente, The Netherlands
Wireless Internet

Jesse M. Heines
University of Massachusetts Lowell
Extensible Stylesheet Language (XSL)

Rodney J. Heisterberg
Notre Dame de Namur University and
Rod Heisterberg Associates
Collaborative Commerce

Steven J. Henry
Wolf, Greenfield & Sacks, P.C.
Open Source Development and Licensing

Julie Hersberger
University of North Carolina at Greensboro
Internet Censorship

Kenneth Einar Himma
University of Washington
Legal, Social, and Ethical Issues

Matthias Holweg
Massachusetts Institute of Technology
Managing the Flow of Materials Across the Supply Chain

Russ Housley
Vigil Security, LLC
Public Key Infrastructure (PKI)

Yeong-Hyeon Hwang
University of Illinois at Urbana–Champaign
Travel and Tourism

Robert E. Irie
SPAWAR Systems Center San Diego
Web Site Design

Linda C. Isenhour
University of Central Florida
Human Resources Management

H.-Arno Jacobsen
University of Toronto, Canada
Application Service Providers (ASPs)

Charles W. Jaeger
Southerrn Oregon University
Cyberterrorism

Dwight Jaffee
University of California, Berkeley
Real Estate

Sushil Jajodia
George Mason University
Intrusion Detection Techniques

Mark Jeffery
Northwestern University
Return on Investment Analysis for E-business Projects

Andrew Johnson
University of Illinois at Chicago
Virtual Reality on the Internet: Collaborative Virtual Reality

Ari Juels
RSA Laboratories
Encryption

Bhushan Kapoor
California State University, Fullerton
ActiveX Data Objects (ADO)

Joseph M. Kayany
Western Michigan University
Internet Etiquette (Netiquette)

Doug Kaye
RDS Strategies LLC
Web Hosting

Chuck Kelley
Excellence In Data, Inc.
Data Warehousing and Data Marts

Diane Ketelhut
Harvard University
Distance Learning (Virtual Learning)

Chang-Su Kim
Seoul National University, Korea
Data Compression

Wooyoung Kim
University of Illinois at Urbana-Champaign
Web Services

Jerry Kindall
Epok Inc.
Digital Identity

Brad Kleindl
Missouri Southern State University–Joplin
Value Chain Analysis

Graham Knight
University College London, United Kingdom
Internet Architecture

Craig D. Knuckles
Lake Forest College
DHTML (Dynamic HyperText Markup Language)

Jim Krause
Indiana University
Enhanced TV

Peter Kroon
Agere Systems
Speech and Audio Compression

Gary J. Krug
Eastern Washington University
Convergence of Data, Sound, and Video

Nir Kshetri
University of North Carolina
Gender and Internet Usage
Global Diffusion of the Internet

C.-C. Jay Kuo
University of Southern California
Data Compression

Stan Kurkovsky
Columbus State University
Common Gateway Interface (CGI) Scripts

Pamela M. H. Kwok
Hong Kong Polytechnic University, China
Wireless Marketing

Jennifer Lagier
Hartnell College
File Types

Thomas D. Lairson
Rollins College
Supply Chain Management and the Internet

Gary LaPoint
Syracuse University
International Supply Chain Management

Haniph A. Latchman
University of Florida
Managing a Network Environment

John LeBaron
University of Massachusetts Lowell
Internet2

Kenneth S. Lee
University of Pennsylvania
Wireless Internet

Jason Leigh
University of Illinois at Chicago
Virtual Reality on the Internet: Collaborative Virtual Reality

Margarita Maria Lenk
Colorado State University
Guidelines for a Comprehensive Security System

Nanette S. Levinson
American University
Developing Nations

Edwin E. Lewis Jr.
Johns Hopkins University
E-business ROI Simulations

David J. Loundy
DePaul University
Online Stalking

Robert H. Lowson
University of East Anglia, United Kingdom
E-systems for the Support of Manufacturing Operations
Supply Networks: Developing and Maintaining Relationships and Strategies

David Lukoff
Saybrook Graduate School and Research Center
Health Issues

Kuber Maharjan
Purdue University
Downloading from the Internet

Julie R. Mariga
Purdue University
Mobile Devices and Protocols
Mobile Operating Systems and Applications

Oge Marques
Florida Atlantic University
Interactive Multimedia on the Web

Prabhaker Mateti
Wright State University
TCP/IP Suite

Bruce R. Maxim
University of Michigan–Dearborn
Game Design: Games for the World Wide Web

Blayne E. Mayfield
Oklahoma State University
Visual C++ (Microsoft)

Cavan McCarthy
Louisiana State University
Digital Libraries

Patrick McDaniel
AT&T Labs
Authentication

David E. McDysan
WorldCom
Virtual Private Networks: Internet Protocol (IP) Based

Daniel J. McFarland
Rowan University
Client/Server Computing

Matthew K. McGowan
Bradley University
Electronic Data Interchange (EDI)

Nenad Medvidovic
University of Southern California
JavaBeans and Software Architecture

Nikunj R. Mehta
University of Southern California
JavaBeans and Software Architecture

John A. Mendonca
Purdue University
Organizational Impact

Weiyi Meng
State University of New York at Binghamton
Web Search Technology

Mark S. Merkow
E-commerce Guide
Secure Electronic Transactions (SET)

Mark Michael
King's College
HTML/XHTML (HyperText Markup Language/ Extensible HyperText Markup Language)
Physical Security

Brent A. Miller
IBM Corporation
Bluetooth^{TM}—A Wireless Personal Area Network

Robert K. Moniot
Fordham University
Software Piracy

Joseph Morabito
Stevens Institute of Technology
Online Analytical Processing (OLAP)

Roy Morris
Capitol College
Voice over Internet Protocol (IP)

Alec Nacamuli
IBM United Kingdom Ltd., United Kingdom
Electronic Funds Transfer

Annette Nellen
San José State University
Public Accounting Firms
Taxation Issues

Dale Nesbary
Oakland University
Nonprofit Organizations

Dat-Dao Nguyen
California State University, Northridge
Business-to-Business (B2B) Internet Business Models

Peng Ning
North Carolina State University
Intrusion Detection Techniques

Mark E. Nissen
Naval Postgraduate School
Intelligent Agents

Won Gyun No
University of Waterloo, Canada
XBRL (Extensible Business Reporting Language): Business Reporting with XML

Eric H. Nyberg
Carnegie Mellon University
Prototyping

Jeff Offutt
George Mason University
Software Design and Implementation in the Web Environment

Donal O'Mahony
University of Dublin, Ireland
Electronic Payment

Robert Oshana
Southern Methodist University
Capacity Planning for Web Services

Dennis O. Owen
Purdue University
Visual Basic

Raymond R. Panko
University of Hawaii at Manoa
Computer Security Incident Response Teams (CSIRTs)
Digital Signatures and Electronic Signatures
Internet Security Standards

Anand Paul
University of Florida
Inventory Management

Thomas L. Pigg
Jackson State Community College
Conducted Communications Media

Paul S. Piper
Western Washington University
Research on the Internet

Benjamin R. Pobanz
Purdue University
Mobile Devices and Protocols

Richard E. Potter
University of Illinois at Chicago
Groupware

Dennis M. Powers
Southern Oregon University
Cyberlaw: The Major Areas, Development, and Provisions

Paul R. Prabhaker
Illinois Institute of Technology
E-marketplaces

Etienne E. Pracht
University of South Florida
Health Insurance and Managed Care

Frederick Pratter
Eastern Oregon University
JavaServer Pages (JSP)

Robert W. Proctor
Purdue University
Human Factors and Ergonomics

Jian Qin
Syracuse University
Web Content Management

Zinovy Radovilsky
California State University, Hayward
Enterprise Resource Planning (ERP)

Jeremy Rasmussen
Sypris Electronics, LLC
Passwords

Peter Raven
Seattle University
Feasibility of Global E-business Projects

Amy W. Ray
Bentley College
Business Plans for E-commerce Projects

Julian J. Ray
Western New England College
Business-to-Business (B2B) Electronic Commerce

Pratap Reddy
Raritan Valley Community College
Internet Navigation (Basics, Services, and Portals)

Drummond Reed
OneName Corporation
Digital Identity

Vladimir V. Riabov
Rivier College
Storage Area Networks (SANs)

Nick Rich
Cardiff Business School, United Kingdom
Managing the Flow of Materials Across the Supply Chain

Malu Roldan
San Jose State University
Marketing Plans for an E-commerce Project

Constantine Roussos
Lynchburg College
JavaScript

Akhil Sahai
Hewlett-Packard Laboratories
Web Services

Eduardo Salas
University of Central Florida
Human Resources Management

Atul A. Salvekar
Intel Corp.
Digital Communication

Pierangela Samarati
Università di Milano, Italy
Disaster Recovery Planning

J. Christopher Sandvig
Western Washington University
Active Server Pages

Robert J. Schalkoff
Clemson University
Rule-Based and Expert Systems

Shannon Schelin
North Carolina State University
E-government

William T. Schiano
Bentley College
Intranets

Roy C. Schmidt
Bradley University
Risk Management in Internet-Based Software Projects

E. Eugene Schultz
University of California–Berkley Lab
Denial of Service Attacks
Windows 2000 Security

Steven D. Schwaitzberg
Tufts-New England Medical Center
Medical Care Delivery

Kathy Schwalbe
Augsburg College
Project Management Techniques

Mark Shacklette
The University of Chicago
Unix Operating System

P. M. Shankar
Drexel University
Propagation Characteristics of Wireless Channels

John Sherry
Purdue University
History of the Internet

Carolyn J. Siccama
University of Massachusetts Lowell
Internet2

Judith C. Simon
The University of Memphis
Java
Law Enforcement
Law Firms

Robert Simon
George Mason University
Middleware

Nirvikar Singh
University of California, Santa Cruz
Digital Economy

Clara L. Sitter
University of Denver
Library Management

Robert Slade
Consultant
Computer Viruses and Worms

Erick D. Slazinski
Purdue University
Structured Query Language (SQL)

Mark Smith
Purdue University
Supply Chain Management Technologies

Lee Sproull
New York University
Online Communities

Charles Steinfield
Michigan State University
Click-and-Brick Electronic Commerce
Electronic Commerce and Electronic Business

Edward A. Stohr
Stevens Institute of Technology
Online Analytical Processing (OLAP)

Dianna L. Stone
University of Central Florida
Human Resources Management

David Stotts
University of North Carolina at Chapel Hill
Perl

Judy Strauss
University of Nevada, Reno
Marketing Communication Strategies

Wayne C. Summers
Columbus State University
Local Area Networks

Jamie S. Switzer
Colorado State University
Virtual Teams

Dale R. Thompson
University of Arkansas
Public Networks

John S. Thompson
University of Colorado at Boulder
Integrated Services Digital Network (ISDN): Narrowband and Broadband Services and Applications

Stephen W. Thorpe
Neumann College
Extranets

Ronald R. Tidd
Central Washington University
Knowledge Management

Herbert Tuttle
The University of Kansas
Video Streaming

Okechukwu C. Ugweje
The University of Akron
Radio Frequency and Wireless Communications

Asoo J. Vakharia
University of Florida
Supply Chain Management

Robert Vaughn
University of Memphis
Law Enforcement

Vasja Vehovar
University of Ljubljana, Slovenia
Benchmarking Internet

Kim-Phuong L. Vu
Purdue University
Human Factors and Ergonomics

Jordan Walters
BCN Associates, Inc.
Managing a Network Environment

Siaw-Peng Wan
Elmhurst College
Online Banking and Beyond: Internet-Related Offerings from U.S. Banks

Youcheng Wang
University of Illinois at Urbana–Champaign
Travel and Tourism

James. L. Wayman
San Jose State University
Biometric Authentication

Scott Webster
Syracuse University
International Supply Chain Management

Jianbin Wei
Wayne State University
Load Balancing on the Internet

Ralph D. Westfall
California State Polytechnic University, Pomona
Telecommuting and Telework

Pamela Whitehouse
Harvard University
Distance Learning (Virtual Learning)

Dave Whitmore
Champlain College
Multiplexing

Russell S. Winer
New York University
Customer Relationship Management on the Web

Raymond Wisman
Indiana University Southeast
Web Search Fundamentals

Paul L. Witt
University of Texas at Arlington
Internet Relay Chat (IRC)

Mary Finley Wolfinbarger
California State University Long Beach
Consumer Behavior

Peter R. Wurman
North Carolina State University
Online Auction Site Management

Cheng-Zhong Xu
Wayne State University
Load Balancing on the Internet

Qiang Yang
Hong Kong University of Science and Technology, China
Machine Learning and Data Mining on the Web

A. Neil Yerkey
University at Buffalo
Databases on the Web

Clement Yu
University of Illinois at Chicago
Web Search Technology

Daniel Dajun Zeng
University of Arizona
Intelligent Agents

Yan-Qing Zhang
Georgia State University
Fuzzy Logic

Xiaobo Zhou
University of Colorado at Colorado Springs
Load Balancing on the Internet

Donald E. Zimmerman
Colorado State University
Usability Testing: An Evaluation Process for Internet Communications

Preface

The Internet Encyclopedia is the first comprehensive examination of the core topics in the Internet field. *The Internet Encyclopedia*, a three-volume reference work with 205 chapters and more than 2,600 pages, provides comprehensive coverage of the Internet as a business tool, IT platform, and communications and commerce medium. The audience includes the libraries of two-year and four-year colleges and universities with MIS, IT, IS, data processing, computer science, and business departments; public and private libraries; and corporate libraries throughout the world. It is the only comprehensive source for reference material for educators and practitioners in the Internet field.

Education, libraries, health, medical, biotechnology, military, law enforcement, accounting, law, justice, manufacturing, financial services, insurance, communications, transportation, aerospace, energy, and utilities are among the fields and industries expected to become increasingly dependent upon the Internet and Web technologies. Companies in these areas are actively researching the many issues surrounding the design, utilization, and implementation of these technologies.

This definitive three-volume encyclopedia offers coverage of both established and cutting-edge theories and developments of the Internet as a technical tool and business/communications medium. The encyclopedia contains chapters from global experts in academia and industry. It offers the following unique features:

1) Each chapter follows a format which includes title and author, chapter outline, introduction, body, conclusion, glossary, cross references, and references. This unique format enables the readers to pick and choose among various sections of a chapter. It also creates consistency throughout the entire series.
2) The encyclopedia has been written by more than 240 experts and reviewed by more than 840 academics and practitioners chosen from around the world. This diverse collection of expertise has created the most definitive coverage of established and cutting edge theories and applications in this fast-growing field.
3) Each chapter has been rigorously peer reviewed. This review process assures the accuracy and completeness of each topic.
4) Each chapter provides extensive online and offline references for additional readings. This will enable readers to further enrich their understanding of a given topic.
5) More than 1,000 illustrations and tables throughout the series highlight complex topics and assist further understanding.
6) Each chapter provides extensive cross references. This helps the readers identify other chapters within the encyclopedia related to a particular topic, which provides a one-stop knowledge base for a given topic.
7) More than 2,500 glossary items define new terms and buzzwords throughout the series, which assists readers in understanding concepts and applications.
8) The encyclopedia includes a complete table of contents and index sections for easy access to various parts of the series.
9) The series emphasizes both technical and managerial issues. This approach provides researchers, educators, students, and practitioners with a balanced understanding of the topics and the necessary background to deal with problems related to Internet-based systems design, implementation, utilization, and management.
10) The series has been designed based on the current core course materials in several leading universities around the world and current practices in leading computer- and Internet-related corporations. This format should appeal to a diverse group of educators, practitioners, and researchers in the Internet field.

We chose to concentrate on fields and supporting technologies that have widespread applications in the academic and business worlds. To develop this encyclopedia, we carefully reviewed current academic research in the Internet field at leading universities and research institutions around the world. Management information systems, decision support systems (DSS), supply chain management, electronic commence, network design and management, and computer information systems (CIS) curricula recommended by the Association of Information Technology Professionals (AITP) and the Association for Computing Management (ACM) were carefully investigated. We also researched the current practices in the Internet field used by leading IT corporations. Our work enabled us to define the boundaries and contents of this project.

TOPIC CATEGORIES

Based on our research we identified 11 major topic areas for the encyclopedia:

- Foundation;
- Infrastructure;
- Legal, social, organizational, international, and taxation issues;
- Security issues and measures;
- Web design and programming;
- Design, implementation, and management;
- Electronic commerce;
- Marketing and advertising on the Web;

- Supply chain management;
- Wireless Internet and e-commerce; and
- Applications.

Although these 11 categories of topics are interrelated, each addresses one major dimension of the Internet-related fields. The chapters in each category are also interrelated and complementary, enabling readers to compare, contrast, and draw conclusions that might not otherwise be possible.

Although the entries have been arranged alphabetically, the light they shed knows no bounds. The encyclopedia provides unmatched coverage of fundamental topics and issues for successful design, implementation, and utilization of Internet-based systems. Its chapters can serve as material for a wide spectrum of courses, such as the following:

- Web technology fundamentals;
- E-commerce;
- Security issues and measures for computers, networks, and online transactions;
- Legal, social, organizational, and taxation issues raised by the Internet and Web technology;
- Wireless Internet and e-commerce;
- Supply chain management;
- Web design and programming;
- Marketing and advertising on the Web; and
- The Internet and electronic commerce applications.

Successful design, implementation, and utilization of Internet-based systems require a thorough knowledge of several technologies, theories, and supporting disciplines. Internet and Web technologies researchers and practitioners have had to consult many resources to find answers. Some of these sources concentrate on technologies and infrastructures, some on social and legal issues, and some on applications of Internet-based systems. This encyclopedia provides all of this relevant information in a comprehensive three-volume set with a lively format.

Each volume incorporates core Internet topics, practical applications, and coverage of the emerging issues in the Internet and Web technologies field. Written by scholars and practitioners from around the world, the chapters fall into the 11 major subject areas mentioned previously.

Foundation

Chapters in this group examine a broad range of topics. Theories and concepts that have a direct or indirect effect on the understanding, role, and the impact of the Internet in public and private organizations are presented. They also highlight some of the current issues in the Internet field. These articles explore historical issues and basic concepts as well as economic and value chain concepts. They address fundamentals of Web-based systems as well as Web search issues and technologies. As a group they provide a solid foundation for the study of the Internet and Web-based systems.

Infrastructure

Chapters in this group explore the hardware, software, operating systems, standards, protocols, network systems, and technologies used for design and implementation of the Internet and Web-based systems. Thorough discussions of TCP/IP, compression technologies, and various types of networks systems including LANs, MANS, and WANs are presented.

Legal, Social, Organizational, International, and Taxation Issues

These chapters look at important issues (positive and negative) in the Internet field. The coverage includes copyright, patent and trademark laws, privacy and ethical issues, and various types of cyberthreats from hackers and computer criminals. They also investigate international and taxation issues, organizational issues, and social issues of the Internet and Web-based systems.

Security Issues and Measures

Chapters in this group provide a comprehensive discussion of security issues, threats, and measures for computers, network systems, and online transactions. These chapters collectively identify major vulnerabilities and then provide suggestions and solutions that could significantly enhance the security of computer networks and online transactions.

Web Design and Programming

The chapters in this group review major programming languages, concepts, and techniques used for designing programs, Web sites, and virtual storefronts in the e-commerce environment. They also discuss tools and techniques for Web content management.

Design, Implementation, and Management

The chapters in this group address a host of issues, concepts, theories and techniques that are used for design, implementation, and management of the Internet and Web-based systems. These chapters address conceptual issues, fundamentals, and cost benefits and returns on investment for Internet and e-business projects. They also present project management and control tools and techniques for the management of Internet and Web-based systems.

Electronic Commerce

These chapters present a thorough discussion of electronic commerce fundamentals, taxonomies, and applications. They also discuss supporting technologies and applications of e-commerce inclining intranets, extranets, online auctions, and Web services. These chapters clearly demonstrate the successful applications of the Internet and Web technologies in private and public sectors.

Marketing and Advertising on the Web

The chapters in this group explore concepts, theories, and technologies used for effective marketing and advertising on the Web. These chapters examine both qualitative and quantitative techniques. They also investigate the emerging technologies for mass personalization and customization in the Web environment.

Supply Chain Management

The chapters in this group discuss the fundamentals concepts and theories of value chain and supply chain management. The chapters examine the major role that the Internet and Web technologies play in an efficient and effective supply chain management program.

Wireless Internet and E-commerce

These chapters look at the fundamental concepts and technologies of wireless networks and wireless computing as they relate to the Internet and e-commerce operations. They also discuss mobile commerce and wireless marketing as two of the growing fields within the e-commerce environment.

Applications

The Internet and Web-based systems are everywhere. In most cases they have improved the efficiency and effectiveness of managers and decision makers. Chapters in this group highlight applications of the Internet in several fields, such as accounting, manufacturing, education, and human resources management, and their unique applications in a broad section of the service industries including law, law enforcement, medical delivery, health insurance and managed care, library management, nonprofit organizations, banking, online communities, dispute resolution, news services, public relations, publishing, religion, politics, and real estate. Although these disciplines are different in scope, they all utilize the Internet to improve productivity and in many cases to increase customer service in a dynamic business environment.

Specialists have written the collection for experienced and not-so-experienced readers. It is to these contributors that I am especially grateful. This remarkable collection of scholars and practitioners has distilled their knowledge into a fascinating and enlightening one-stop knowledge base in Internet-based systems that “talk” to readers. This has been a massive effort but one of the most rewarding experiences I have ever undertaken. So many people have played a role that it is difficult to know where to begin.

I should like to thank the members of the editorial board for participating in the project and for their expert advice on the selection of topics, recommendations for authors, and review of the materials. Many thanks to the more than 840 reviewers who devoted their times by proving advice to me and the authors on improving the coverage, accuracy, and comprehensiveness of these materials.

I thank my senior editor at John Wiley & Sons, Matthew Holt, who initiated the idea of the encyclopedia back in spring of 2001. Through a dozen drafts and many reviews, the project got off the ground and then was managed flawlessly by Matthew and his professional team. Matthew and his team made many recommendations for keeping the project focused and maintaining its lively coverage. Tamara Hummel, our superb editorial coordinator, exchanged several hundred e-mail messages with me and many of our authors to keep the project on schedule. I am grateful to all her support. When it came to the production phase, the superb Wiley production team took over. Particularly I want to thank Deborah DeBlasi, our senior production editor at John Wiley & Sons, and Nancy J. Hulan, our project manager at TechBooks. I am grateful to all their hard work.

Last, but not least, I want to thank my wonderful wife Nooshin and my two lovely children Mohsen and Morvareed for being so patient during this venture. They provided a pleasant environment that expedited the completion of this project. Nooshin was also a great help in designing and maintaining the author and reviewer databases. Her efforts are greatly appreciated. Also, my two sisters Azam and Akram provided moral support throughout my life. To this family, any expression of thanks is insufficient.

Hossein Bidgoli
California State University, Bakersfield

Guide to the Internet Encyclopedia

The Internet Encyclopedia is a comprehensive summary of the relatively new and very important field of the Internet. This reference work consists of three separate volumes and 205 chapters on various aspects of this field. Each chapter in the encyclopedia provides a comprehensive overview of the selected topic intended to inform a board spectrum of readers ranging from computer professionals and academicians to students to the general business community.

In order that you, the reader, will derive the greatest possible benefit from *The Internet Encyclopedia,* we have provided this Guide. It explains how the information within the encyclopedia can be located.

ORGANIZATION

The Internet Encyclopedia is organized to provide maximum ease of use for its readers. All of the chapters are arranged in alphabetical sequence by title. Chapters titles that begin with the letters A to F are in Volume 1, chapter titles from G to O are in Volume 2, and chapter titles from P to Z are in Volume 3. So that they can be easily located, chapter titles generally begin with the key word or phrase indicating the topic, with any descriptive terms following. For example, "Virtual Reality on the Internet: Collaborative Virtual Reality" is the chapter title rather than "Collaborative Virtual Reality."

Table of Contents

A complete table of contents for the entire encyclopedia appears in the front of each volume. This list of titles represents topics that have been carefully selected by the editor-in-chief, Dr. Hossein Bidgoli, and his colleagues on the Editorial Board.

Following this list of chapters by title is a second complete list, in which the chapters are grouped according to subject area. The encyclopedia provides coverage of 11 specific subject areas, such as E-commerce and Supply Chain Management. Please see the Preface for a more detailed description of these subject areas.

Index

The Subject Index is located at the end of Volume 3. This index is the most convenient way to locate a desired topic within the encyclopedia. The subjects in the index are listed alphabetically and indicate the volume and page number where information on this topic can be found.

Chapters

Each chapter in *The Internet Encyclopedia* begins on a new page, so that the reader may quickly locate it. The author's name and affiliation are displayed at the beginning of the article.

All chapters in the encyclopedia are organized according to a standard format, as follows:

- Title and author,
- Outline,
- Introduction,
- Body,
- Conclusion,
- Glossary,
- Cross References, and
- References.

Outline

Each chapter begins with an outline indicating the content to come. This outline provides a brief overview of the chapter so that the reader can get a sense of the information contained there without having to leaf through the pages. It also serves to highlight important subtopics that will be discussed within the chapter. For example, the chapter "Computer Literacy" includes sections entitled Defining a Computer, Categories of Computers According to Their Power, and Classes of Data Processing Systems. The outline is intended as an overview and thus lists only the major headings of the chapter. In addition, lower-level headings will be found within the chapter.

Introduction

The text of each chapter begins with an introductory section that defines the topic under discussion and summarizes the content. By reading this section the readers get a general idea about the content of a specific chapter.

Body

The body of each chapter discusses the items that were listed in the outline section.

Conclusion

The conclusion section provides a summary of the materials discussed in each chapter. This section imparts to the readers the most important issues and concepts discussed within each chapter.

Glossary

The glossary contains terms that are important to an understanding of the chapter and that may be unfamiliar to the reader. Each term is defined in the context of the particular chapter in which it is used. Thus the same term may be defined in two or more chapters with the detail of the definition varying slightly from one to another. The encyclopedia includes approximately 2,500 glossary terms.

For example, the article “Computer Literacy” includes the following glossary entries:

Computer A machine that accepts data as input, processes the data without human interference using a set of stored instructions, and outputs information. Instructions are step-by-step directions given to a computer for performing specific tasks.

Computer generations Different classes of computer technology identified by a distinct architecture and technology; the first generation was vacuum tubes, the second transistors, the third integrated circuits, the fourth very-large-scale integration, and the fifth gallium arsenide and parallel processing.

Cross References

All the chapters in the encyclopedia have cross references to other chapters. These appear at the end of the chapter, following the text and preceding the references. The cross references indicate related chapters which can be consulted for further information on the same topic. The encyclopedia contains more than 2,000 cross references in all. For example, the chapter “Java” has the following cross references:

JavaBeans and Software Architecture; Software Design and Implementation in the Web Environment.

References

The reference section appears as the last element in a chapter. It lists recent secondary sources to aid the reader in locating more detailed or technical information. Review articles and research papers that are important to an understanding of the topic are also listed. The references in this encyclopedia are for the benefit of the reader, to provide direction for further research on the given topic. Thus they typically consist of one to two dozen entries. They are not intended to represent a complete listing of all materials consulted by the author in preparing the chapter. In addition, some chapters contain a Further Reading section, which includes additional sources readers may wish to consult.

A

Active Server Pages

J. Christopher Sandvig, *Western Washington University*

INTRODUCTION

Active server pages (ASP) and ASP.NET are server-side programming technologies developed by Microsoft Corporation. Server-side programs run on Web servers and are used for dynamically generating Web pages. Many data-intensive Web applications, such as Web-based e-mail, online banking, news, weather, and search engines, require the use of server-side programs. Server-side programming is also useful for many other applications, such as collecting data, processing online payments, and controlling access to information. Server-side programs are executed each time a Web user requests a dynamically generated Web page. ASP and ASP.NET Web pages are identified by the file extensions .asp and .aspx, respectively.

The primary advantage of server-side programming technologies is their ability to utilize databases. Databases are very efficient and powerful tools used for storing and retrieving data. Most sophisticated Web applications utilize databases for storing their data. Server-side programming is also very reliable. Servers provide a more stable, secure, and controllable programming environment then the client's browser.

In addition to ASP and ASP.NET other popular server-side Web technologies include Perl, PHP, J2EE, Java server pages, Python, and ColdFusion. All of these technologies run on the Web server and generate their output in HTML (hypertext markup language). The Web server sends the HTML to the client's browser, which interprets it and displays it as formatted text. Most server-side technologies have similar capabilities; the primary functional differences between them are scalability and programming complexity. These issues are discussed later in the chapter.

Server-side technologies may work in conjunction with client-side technologies. Client-side technologies are executed by the user's browser and include JavaScript, Java applets, and ActiveX controls. Client-side technologies are typically used for controlling browser display features, such as mouse rollovers, dynamic images, and opening new browser windows.

The advantage of client-side technologies is that the processing is done on the client's machine, thus reducing the load on the Web server. Client-side technologies can execute more quickly because they do not require the back-and-forth transmission of data between the client and the server. However, the functionality of client-side technologies is limited due to their inability to access databases. They also require sending more data to the client, which can increase the time required to load a Web page.

Server-side technologies are more reliable than client-side technologies. Because client-side technologies run on the user's browser, they are dependent on the capabilities of the browser. Because browsers vary in their capabilities, client-side code that works well on one browser may not work on another. Server-side technologies, on the other hand, are always executed on the server. Because developers know the capabilities of their server, the results are very predictable. They send only HTML to the client's browser, which all browsers support fairly consistently.

Introduction of ASP and ASP.NET

Microsoft introduced ASP in December 1997 as a feature of its Internet Information Server (IIS) 3.0. Most of the other popular server-side technologies were introduced prior to 1995, making ASP a relatively late entrant into the world of server-side technologies. One year later, in

December 1998, Microsoft released ASP 2.0 as part of the Windows NT4 option pack (a free download). Two years later, IIS 3.0 was introduced as part of Windows 2000. Each release introduced modest improvements in both functionality and performance.

Despite its late introduction, ASP quickly became a popular server-side technology. Its popularity was driven by both its ease of programming and the widespread availability of the IIS server, which is bundled free with several versions of Microsoft's Windows operating system. Major Web sites that use ASP include Barnes and Noble (http://www.bn.com), Dell Computer (http://www.dell.com), JC Penney (http://www.jcpenney.com), MSNBC (http://www.msnbc.com), Ask.com (http://www.ask.com), and Radio Shack (http://www.radioshack.com).

ASP.NET 1.0 was introduced by Microsoft in February 2002. ASP.NET differs significantly from ASP in its syntax, performance, and functionality. In many ways ASP.NET is a new product rather than simply an upgrade of ASP. ASP.NET was designed to support Microsoft's .NET strategy and has extensive support for two technologies that are at the core of the strategy: XML (extensible markup language) and SOAP (simple object access protocol). It was also designed to overcome some of the weakness of ASP in the area of scalability and reliability. The differences between ASP and ASP.NET are discussed in greater detail later in this chapter.

At the time of this writing, ASP.NET had been just recently introduced and had not yet been adopted by any major Webs sites. However, it has received many positive reviews from developers and is expected to capture 30% of the enterprise development market by 2004 (Sholler, 2002).

ASP and ASP.NET are both supported by Microsoft's IIS server. ASP is supported by all versions of IIS from 3.0 up and later. ASP engines are also available from third-party developers that support ASP on a number of non-Microsoft operating systems and Web servers. Vendors include ChilliSoft and Stryon.

The ASP.NET framework is a free download from Microsoft and runs with IIS under Windows 2000 and later. It is designed for portability between operating systems, and it is expected that third-party vendors will provide ASP.NET compilers for non-Microsoft servers and operating systems.

Framework

Both ASP and ASP.NET are programming frameworks rather than programming languages. A framework is a bundle of technologies that work together to provide the tools needed for creating dynamic Web pages.

The ASP framework provides support for two scripting languages, seven server objects, and Microsoft's ActiveX data objects (ADO). The two scripting languages supported by Microsoft's IIS server are VBScript and Jscript, of which VBScript is the most popular because of its similarity to the widely used Visual Basic programming language. Third-party vendors offer ASP scripting engines that support other scripting languages, such as Perl and Python.

Much of ASP's functionality is derived from a collection of seven server objects. These intrinsic objects provide the tools for sending output to the user's browser, receiving input from the client, accessing server resources, storing data, writing files, and many other useful capabilities. The seven objects are application, ASPerror, objectcontext, request, response, session, and server. Each object has a set of properties, methods, and events that are employed by the ASP programmer to access the object's functionality. A detailed description of ASP's server objects is beyond the scope of this chapter, but an excellent reference is available from Microsoft's developer library: http: // msdn.microsoft.com / library / default.asp ? URL=/ library/psdk/iisref/vbob74bw.htm.

The ASP framework provides database access through the recordset object, which is a member of Microsoft's ActiveX data objects (ADO). The recordset allows ASP scripts to utilize most commercial database products, including Microsoft Access, Microsoft SQL Server, Informix, and Oracle (Mitchell and Atkinson, 2000).

The ASP.NET framework supports a large number of programming languages and operating systems. Microsoft's .NET framework provides native support for VB.NET, C# (pronounced C sharp), and Jscript.NET. Third-party vendors have announced plans to produce ".NET-compliant" compilers for over a dozen other languages, including Eiffel, Pascal, Python, C++, COBOL, Perl, and Java (Kiely, 2002; Ullman, Ollie, Libre, & Go, 2001).

The ASP.NET framework replaces ASP's seven server objects with an extensive "base class library." The class library contains hundreds of classes and offers considerably more functionality than do ASP's seven server objects. ASP.NET programmers access this code by initiating objects from the class library. All ASP.NET programs utilize the same base class library regardless of the programming language used.

Scripting Versus Object-Oriented Programming Languages

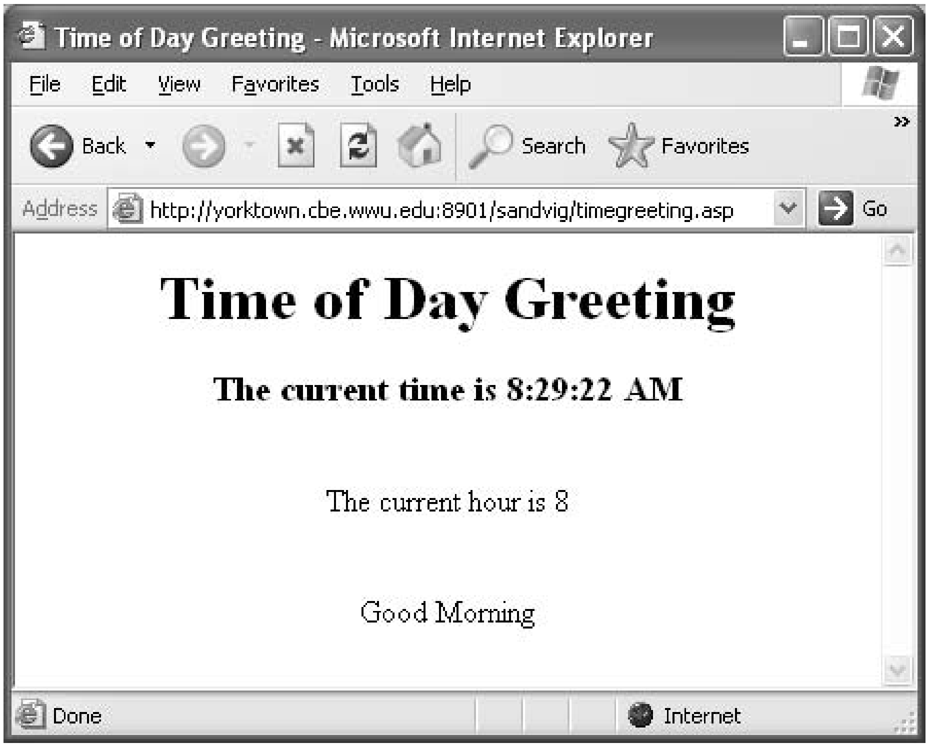
An important difference between ASP and ASP.NET is that ASP uses scripting languages whereas ASP.NET supports object-oriented programming languages. Scripting languages are generally easier to write and understand, but their simple structure does not lend itself well to complex programs. Scripting languages are usually interpreted languages, meaning that the server compiles them each time a page is served.

ASP.NET supports object-oriented event-driven programming languages. Object-oriented languages organize computer code into small units of functionality, called objects. The advantage of these languages is that they encapsulate functionality into small reusable objects. Writing and understanding object-oriented programs can be more difficult than understanding scripts, but the encapsulation of program functionality into discrete, reusable objects offers considerable advantages in complex programs.

CODE EXAMPLES

ASP

The following two code samples illustrate how ASP and ASP.NET work. The ASP code in Listing 1 displays a time-of-day message. The scripting language used is


Time of Day Greeting - Microsoft Internet Explorer
File
Edit
View
Favorites
Tools
Help
Back
Search
Favorites
Address
http://yorktown.cbe.wwu.edu:8901/sandvig/timegreeting.asp
Go
Done
Internet

Time of Day Greeting

The current time is 8:29:22 AM

The current hour is 8

Good Morning

Figure 1: ASP example: code Listing 1 viewed through a Web browser.

VBScript and the file is saved on the Web server as TimeGreeting.asp. The code enclosed by the <% and%> tags is executed on the server and is a mixture of VBScript and ASP server objects. The code inside the < > tags is HTML, which describes how the output should be formatted on the user's browser.

A close look at Listing 1 illustrates how ASP works. The first tag, `<% @LANGUAGE=VBScript %>`, tells the server which scripting language is used. Because VBScript is ASP's default language, this tag is not required, but it is considered good programming practice to include it.

The second set of server tags, `<% response.write Time()%>`, instructs the server to insert the current time into the output. Time() is a VBScript function that returns the current time of day from the server's internal clock. The response object's write method is used to direct the output to the user's browser.

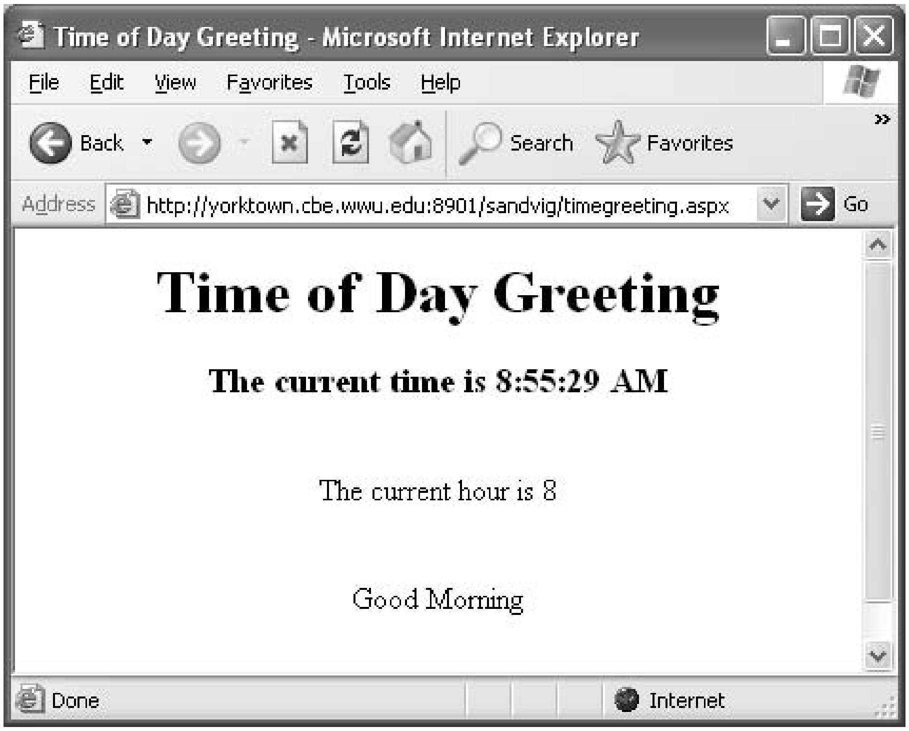
The third set of tags, `<% response.write hour (time())%>`, is similar to the previous line but the time function is nested inside the VBScript hour() function. The hour() function strips the minutes and seconds from the time and returns only the hour portion of the time as an integer value between 0 and 23.

The fourth set of tags uses an "If then . . . End If" statement to send an appropriate time-of-day message. The If then . . . End If statement evaluates expressions as either true or false. If an expression is true, then the statements that immediately follow it are executed, otherwise execution jumps to the next conditional statement (ElseIf). Each successive statement is evaluated until the first true one is found. If none of the conditions is true, then the statements following the "Else" statement are executed. If more than one expression is true, then only the first is evaluated and its associated statements are executed. Figure 1 shows the output of this script displayed on a Web browser.

Listing 1: Source code for TimeGreeting.asp.

```
<% @LANGUAGE=VBScript %>
<HTML>
<HEAD>
   <TITLE>Time of Day Greeting</TITLE>
</HEAD>
<BODY>
   <center>
      <h1>Time of Day Greeting</h1>
      <h3>The current time is
      <% response.write Time()%></h3><br>
      <p>The current hour is
         <% response.write hour(time())%>
      </p><br>
<%     IF Hour(time) < 12 then
          response.write "Good Morning"
       ElseIf hour(time) < 18 then
          response.write "Good Afternoon"
       ElseIf hour(time) < 22 then
          response.write "Good Evening"
       ELSE
          response.write "Good Night"
       END IF
%>
</center>
</body></html>
```

Listing 2 shows the HTML output generated by the ASP code in Listing 1. (This output was obtained by viewing the page on a browser, clicking on the right mouse button, and selecting "View Source.") Note that all of



Time of Day Greeting - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address
http://yorktown.cbe.wwu.edu:8901/sandvig/timegreeting.aspx
Go
Done
Internet

Time of Day Greeting

The current time is 8:55:29 AM

The current hour is 8

Good Morning

Figure 2: ASP.NET example: code Listing 3 viewed through a Web browser.

the code inside the <%%> tags has been processed by the server and replaced with the appropriate output. The client can view the HTML output that results from executing the server-side code enclosed within the <%%> tags but cannot view the server-side code that generated the output.

Listing 2: HTML sent to the browser after the server has processed TimeGreeting.asp.

```
<HTML>
<HEAD>
     <TITLE>Time of Day Greeting</TITLE>
</HEAD>
<BODY>
     <center>
         <h1>Time of Day Greeting</h1>
         <h3>The current time is
           8:29:22 AM</h3><br>
         <p>The current hour is 8</p><br>
         Good Morning
   </center>
</body>
</html>
```

ASP.NET

ASP.NET code differs from ASP code in many important ways. The most visible difference is the separation of the server-side code and the HTML. This is illustrated in Listing 3. The server-side code is located at the top of the page within <script> tags and the HTML is located below it. ASP.NET uses "Web controls" to insert the output of the server-side code into the HTML. The Web controls are themselves objects with properties, methods, and events.

The first line of code in Listing 3 contains a <script> tag. This tag instructs the compiler that the programming language is VB.NET and that the code enclosed within the <script> tags should be executed on the server.

The second line of Listing 3 defines a subroutine named Page_Load(). The Page_Load subroutine runs automatically each time the page is loaded. The first line within the Page_Load subroutine,

```
lbTime.text = DateTime.Now.ToString("T")
```

obtains the current time from the system clock and formats it as a string. The DateTime object is a member of ASP.NET's base class library. The current time obtained from the DateTime object is assigned to the text property of a label named lbTime. The label object is instantiated and named in the HTML portion of the page with the statement

```
<asp:label id=lbTime runat="server"/>
```

The current hour and time greeting are assigned to the label controls named lbHour and lbGreeting, respectively. This output is inserted into the HTML as before. The output produced by TimeGreeting.aspx is shown in Figure 2. Note that despite the differences in the code, the output is same as that produced by the ASP code shown in Listing 1. Underscore (_) indicates continuation of a line.

Listing 3: ASP.NET source code for TimeGreeting.aspx.

```
<script language="vb" runat="server">
   Sub Page_Load()
      lbTime.text = _
         DateTime.Now.ToString("T")
      lbHour.text = _
         hour(DateTime.Now.ToString("T"))
      If hour(DateTime.Now. _
        ToString("T"))<12 then
        lbGreet.text = "Good Morning"
      ElseIf hour(DateTime.Now. _
        ToString("T"))<18 then
        lbGreet.text = "Good Afternoon"
      ElseIf hour(DateTime.Now. _
        ToString("T"))<22 then
        lbGreet.text = "Good Evening"
      ELSE
        lbGreet.text = "Good Night"
      End If
  End Sub
</script>
<html>
<head>
  <title>Time of Day Greeting</title>
</head>
<body>
<center>
  <h1>Time of Day Greeting</h1>
  <h3>The current time is
    <asp:label id=lbTime runat="server"/>
  </h3><br>
  <p>The current hour is
    <asp:label id=lbHour runat="server"/>
  <br>
  <p><asp:label id=lbGreet runat="server"/>
</center>
</body></html>
```

This example illustrates how ASP.NET separates the code from the presentation by placing all the code inside <script> tags and assigning the code output to Web controls. In this example, only one Web control is used, the label control, but ASP.NET's base class library contains dozens of Web controls. Many of them are quite sophisticated and offer a wide variety of options for displaying data.

The example also illustrates how ASP.NET encapsulates code within subroutines. The subroutines are called by events, such as the page being requested by a client or a client clicking a button on the page. Encapsulating code into discrete objects and then calling them when needed is advantageous for complex programs. Contrast this method to ASP, which executes code in a linear fashion starting at the top of the script.

ASP.NET is a central feature of Microsoft's .NET strategy and is expected to play a large role in the future of both Microsoft and the Internet. The following sections of this chapter overview the major features of ASP.NET.

MICROSOFT'S .NET STRATEGY

Microsoft's .NET strategy evolved from a series of e-mail messages sent from Microsoft President Bill Gates to Microsoft employees during the mid-1990s. The central point of the e-mail messages was that the world of computing was expanding from the desktop to the Internet and that Microsoft's new products were going to lead the way. By 2001, this strategic direction had evolved into Microsoft's .NET strategy.

The .NET strategy is articulated in an open letter, dated June 18, 2001, from Bill Gates to all information technology professionals (Gates, 2001). In this letter, Gates observes that the Internet is similar to the old mainframe world in which information is stored in large centralized databases. There is little communication between applications, creating "islands of functionality and data." This "server-centric computing model" results in huge inefficiencies because programs and data are replicated many times on different servers and because the functionality of many applications is diminished by their limited access to information.

A primary objective of the .NET strategy is to create the tools that allow applications to talk to each other. Gates' vision is that programs using the .NET model "will run across multiple Web sites, drawing on information and services from each of them and combining and delivering them in customized form to any device" (Gates, 2001).

The ASP.NET framework is a central feature of Microsoft's .NET strategy. The following sections of this chapter discuss the distinctive features of the ASP.NET framework.

Web Services

Web Services is a name given to a bundle of technologies that allows applications to communicate with each other. Web services create connections between the "islands of functionality and data."

An address validation service illustrates the benefits of Web services. Currently, a merchant wishing to validate customer-supplied shipping addresses needs to maintain a large database of valid addresses. Many merchants incur the high cost of maintaining such databases. Each merchant's database is accessible only to its owner and represents an "island of data."

Web services eliminate the need for merchants to incur the cost of maintaining such databases. Instead they can subscribe to a Web-based "address validation service." When a customer enters a shipping address on the merchant's Web site, the merchant's Web server sends the information to the Web service. The Web service immediately validates the address and returns the results. Invalid addresses are flagged, and the user is asked to correct it before his or her order is accepted. The service can even send a list of valid alternatives.

The entire transaction can be completed between the merchant's computer and the service provider's computer in a fraction of a second. The Web service can be shared by thousands of merchants, eliminating the cost of maintaining multiple databases.

Web services are already becoming available. United Parcel Service (UPS) offers several Web services, including address validation, shipping rate calculations, and package tracking. UPS provides these services as a free service to its customers. Providing Web services to its customers saves UPS money because fewer packages are shipped to incorrect address. They are also very convenient for UPS customers, provide them with an incentive to ship via UPS. More information on the Web services offered by UPS is available at http://www.ec.ups.com/ecommerce/solutions/c1.html.

ASP.NET's base class library provides many classes for supporting Web services. Microsoft hopes that ASP.NET's powerful tools for deploying and consuming Web services will provide it with an entrée into the profitable high-end corporate software market (Buckman & Bulkeley, 2002).

XML and SOAP

ASP.NET provides extensive support for the two key technologies underlying Web services: XML (extensible markup language) and SOAP (simple object access protocol). XML is similar to HTML in the sense that both are used to describe text. The difference is that HTML formats text for human consumption whereas XML formats text for machine consumption. HTML defines how data is displayed on a page, specifying, for instance, the fonts, font sizes, font color, and layout of the text used to display the data. XML describes the content of the data. For example, if the data consists of a book title and author, XML will label the data <title> and <author>, respectively. SOAP defines the protocols used for passing XML documents between computers. Both XML and SOAP are open industry standards managed by the World Wide Web Consortium (http://www. W3C.org). More information about XML and SOAP is available from the Organization for the Advancement of Structured Information Systems (http://www.xml.org).

Compiled Code

An important difference between ASP and ASP.NET is that ASP uses interpreted languages whereas ASP.NET uses compiled languages. Interpreted languages are translated from human-readable source code to machine-readable compiled code each time a program is called. This has the advantage of simplicity, because only one copy of the program needs to be stored on the computer. A major disadvantage of interpreted languages is that frequent compiling can cause performance problems on high-volume sites.

Compiled languages, such as those used by ASP.NET, are compiled once and stored in a compiled form. This eliminates the need to recompile the code each time the page is served. ASP.NET stores the compiled code as Microsoft intermediate language (MSIL), which is portable between operating systems. Each time an ASP.NET page is called, the MSIL code undergoes a final stage of compilation, called JIT (just in time) compilation. ASP.NET's compiled code is considerably more efficient than ASP's interpreted code, resulting in a significant increase in performance.

Another advantage of compiled code is that it can protect code from being plagiarized. Software companies make large investments in developing computer applications and usually do not want others to view their source code. Applications are typically distributed in a compiled format so that they will execute on computers but are unreadable to potential thieves. ASP scripts can be compiled, but the resulting components must go through an installation process on the server before they may be executed.

Common Language Run Time

The common language run time (CLR) manages the execution of ASP.NET code. The CLR checks to make sure that the MSIL reflects the most recent version of the source code. If the source code has been updated, the CLR recompiles the source code to MSIL before serving the page to the client. The CLR also provides such services as error handling, security features, and cross-language integration (Sussman, 2001; Payne, 2002). Code that is managed by the CLR is called managed code. The automated management features provided by the CLR saves ASP.NET developers the time and effort needed to manage their code manually.

Language and Platform Independence

An important feature of MSIL is that it provides both language and platform independence. An ASP.NET page can be written in any one of several CLR-supported programming languages. MSIL is independent of the programming language in which the source code is written.

MSIL can be compiled to run on any operating system for which a .NET-compliant JIT compiler is available. Microsoft's ASP.NET framework supports only Windows operating systems (9x, NT, 2000, and XP), but it is widely expected that third-party vendors and open development initiatives will produce JIT compilers for a variety of operating systems, including the popular UNIX and Linux operating systems. An open development initiative named the Mono Project has developed an open source, Unix version of the Microsoft .NET development platform (Mono Project).

Separation of Code and Content

Listing 3 illustrates how server-side code is separated from the HTML. The advantage of this structure is that the code is cleaner and easier to read. This division of code and output eliminates the infamous "spaghetti code" that can occur when program logic and formatting logic are intermingled.

Support for Multiple Client Types

The delivery of Web content is expanding to a broad range of digital devices, including personal digital assistants (PDAs), cell phones, and Internet appliances. The proliferation of Web-capable output devices makes it increas ingly important that server-side technologies be able to recognize the output device and format their output appropriately. ASP.NET automatically checks the client's output device and modifies the format of its output to match the device. This eliminates the need for developers to customize each application to handle different output devices, thereby reducing the cost of application development.

Modularity

ASP.NET uses event-driven object-oriented programming languages. One of the primary advantages of object-oriented languages is that they encapsulate program functionality into discrete reusable modules. Such modules can be written once, thoroughly tested, and reused many times. Because modules are self-contained and have well-defined inputs and outputs, they have the effect of reducing overall program complexity. Well-planned and programmed modules are easier to debug and more robust than non-object-oriented programs. The benefits of modularity are especially pronounced in large, complex applications.

Base Class Library

ASP.NET derives much of its functionality from an enormous library of prewritten code called the base class library. The library is organized into a hierarchal structure of *namespaces* and classes. Namespaces contain groupings of related classes. For instance, the Namespace System.Web.UI.WebControls contains over 90 classes that provide Web user interface controls. The label control used in Listing 3 is a member of the WebControl namespace, as are controls for creating text boxes, check boxes, radio buttons, and many other user interface controls. The .NET framework contains nearly 100 namespaces and provides a tremendous amount of functionality. Microsoft's base class library documentation may be viewed at the Microsoft developer's library (http://msdn.microsoft.com/library/default.asp?url = /library/en-us/cpref/html/cpref-start.asp).

Session-State Management

One limitation faced by Internet applications is that the client and the server communicate only intermittently. Communication occurs only when the client is sending a request to the server or when the server is responding. Between these short bursts of activity there is no communication between the client and the server (Bidgoli, 1999). This model of communication was quite satisfactory in the early days of the Internet, when most Web content was static, much like the page in a book. However, as the uses for the Web have expanded into transaction-oriented activities, such as online banking, e-mail, and stock trading, it has become increasingly important that the server be able to keep track of the status of each client. Keeping track of the status of each client's session is known as session-state management.

Online shopping carts are a good example of session-state management. Most online e-commerce sites offer shopping carts that allow customers to keep a list of items that they wish to purchase. This list is known as a *virtual shopping cart*. From the perspective of the server, each client is a session and all the information specific to the clients, such as what they have in their carts, is their session state. To manage the clients' shopping carts, the server must be able to keep track of which clients have added which items to their carts. Session state may also include client-specific information, such as credit card information, passwords, and account numbers. A single Web server may be managing hundreds of clients at one time.

There are several tools that can be used for managing session state including cookies, session objects, and the ASP.NET view state. Cookies are one of the most popular tools used for managing session state. Cookies are small text files that the server writes to the client's hard drive. The information contained within the cookies is sent to the server each time a client sends a request. The server can check the cookie information to identify which client sent the request and to retrieve session state that it has stored in the cookie. Cookies are supported by most server-side programming technologies.

Session objects allow developers to store client-specific information in the server's memory. This is useful when the developer does not want the client to be able to view or change the data. ASP also supports session objects, but it requires that they be used in conjunction with cookies. ASP.NET supports session objects without the use of cookies by embedding a unique session ID into the page URL.

View state is a powerful new tool available to ASP.NET developers and is the subject of the next section of this chapter.

View State

ASP.NET also introduced a new method of maintaining page state, called view state. All the information contained within a Web page's Web controls, such as text boxes, radio buttons, labels, and data grids, is saved in the page view state. ASP.NET automatically stores each page's view state in a hidden field, named _VIEWSTATE.

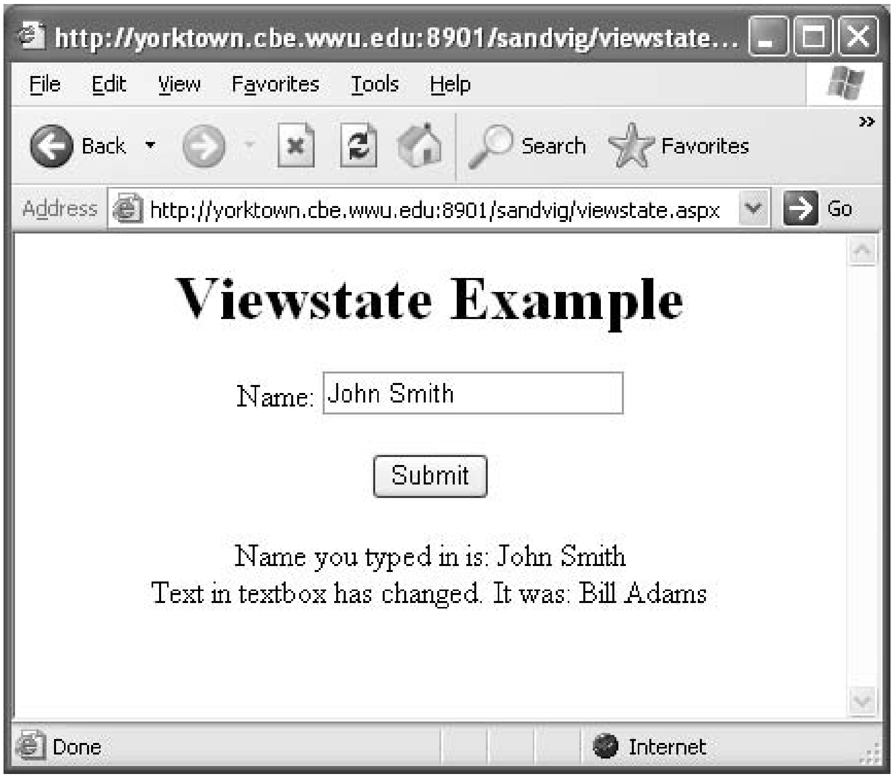
Figure 3 illustrates how view state works. In this example, the user has typed the name "Bill Adams" into the text box and clicked the submit button. He then typed the name "John Smith" into the text box and clicked the submit button again. The ASP.NET program is able to compare the name saved in the view state to the name in the text box and to determine that the user has changed the name in the text box.

Listing 4 shows the server-side source code that produced Figure 3. This page contains four Web controls: a text box, a button, and two labels. The Web controls are created in the HTML portion of the page. The text box and the button are both HTML-form elements and are enclosed within <form> tags. The contents of the Web controls are accessed programmatically within the <script> tabs at the top of the page.

When the user clicks the submit button, the page is submitted to the server. The server automatically executes the code within the Page_Load subroutine. The IF THEN statement then checks to see whether the text in the text box has changed by comparing its current value to its previous value, stored in tbCustomerName.text and ViewState("Name"), respectively. If the text has changed, a message is sent to the user that lists the original text stored in the view state.

The view state saves the data within all the Web controls, such as text boxes, check boxes, radio buttons, and text areas. That is why the name "John Smith" still appears in the text box in Figure 3 even after the user has clicked the submit button and the page has reloaded. Without the ASP.NET's view state feature, the text box would be empty.

Listing 5 show the source code that the ASP.NET has sent to the browser to produce Figure 3. This code shows how ASP.NET has stored the view state in a hidden form field, named "_VIEWSTATE." The view state contains an encrypted version of the information contained in the Web controls when the page was last submitted to the server. View state is useful to developers because it automatically helps maintain many types of session-state information.


http://yorktown.cbe.wwu.edu:8901/sandvig/viewstate...
File Edit View Favorites Tools Help
Back
Search
Favorites
Address
Go
Submit
Done
Internet

http://yorktown.cbe.wwu.edu:8901/sandvig/viewstate.aspx

Viewstate Example

Name: John Smith

Name you typed in is: John Smith
Text in textbox has changed. It was: Bill Adams

Figure 3: View-state example: user has changed text in textbox.

Listing 4: ASP.NET view state example of a server-side source code.

```
<script language="vb" runat="server">
   Sub Page_Load()
      lbMessage1.text = "Name you typed in is: " &  tbCustomerName.text
      If tbCustomerName.text <> ViewState("Name") then
          lbMessage2.text = "Text in textbox has changed. It was: " & _
                             ViewState("Name")
          ViewState("Name") = tbCustomerName.text
      else
          lbMessage2.text = "Name in textbox has NOT changed"
      end if
   End Sub
</script>
<html>
<body>
   <center>
      <h1>Viewstate Example</h1>
      <form runat="server" >
         Name:
         <asp:textbox id=tbCustomerName runat="server"/><br><br>
         <asp:button  id=btSubmit
                      text="Submit"
                      runat="server"/><br><br>
      </form>
      <asp:label id=lbMessage1 runat="server"/><br>
      <asp:label id=lbMessage2 runat="server"/>
   </center>
</body>
</html>
```

Listing 5: ASP.NET view state example of a client-side source code for Figure 3.

```
<html>
<body>
   <center>
      <h1>Viewstate Example</h1>
      <form name="_ctl0" method="post" action="viewstate.aspx" id="_ctl0">
<input type="hidden" name="__VIEWSTATE"
value="dDwxNDg0MDY4OTg0O3Q8cDxsPE5hbWU7PjtsPEpvaG4gU21pdGg7Pj47bDxpPDM+O2k8NT47
PjtsPHQ8cDxwPGw8VGV4dDs+O2w8TmFtZSB5b3UgdHlwZWQgaW4gaXM6IEpvaG4gU21pdGg7Pj47Pjs
7Pjt0PHA8cDxsPFRleHQ7PjtsPFRleHQgaW4gdGV4dGJveCBoYXMgY2hhbmdlZC4gIEl0IHdhczogQm
lsbCBBZGFtczs+Pjs+Ozs+Oz4+Oz5gY/hOuOjX7bLkR2EIrWhP915qQw==" />
             Name:
             <input name="tbCustomerName" type="text" value="John Smith"
                    id="tbCustomerName" /><br><br>
             <label for="cbSendInfo">Send More Information</label><br><br>
             <input type="submit" name="btSubmit" value="Submit"
                    id="btSubmit" /><br><br>
       </form>
       <span id="lbMessage1">Name you typed in is: John Smith</span><br>
       <span id="lbMessage2">Text in textbox has changed. It
            was: Bill Adams</span>
   </center>
</body>
</html>
```

Scalability and Reliability

ASP.NET is designed to meet the needs of high-volume, mission-critical applications. Such applications require software that is both scalable and reliable. A scalable application is one that can handle high transaction volumes. ASP.NET provides scalability through its use of compiled code, which reduces its demands on the server's central processing unit (CPU). Another important scalability feature is its ability to maintain session states across both Web farms (multiple servers) and Web gardens (single servers with multiple CPUs).

ASP.NET's scalability is also enhanced through its support for page caching. Caching is the ability to save the output from running a program and then reusing the saved output on subsequent requests without the need to rerun the underlying code. Caching can dramatically increase the efficiency of serving pages that are frequently viewed but whose content changes infrequently. For instance, a news Web site that is viewed thousands of times per day but is only updated once a day would be a good candidate for caching. ASP.NET allows caching of entire pages or portions of pages.

Reliability is an important performance criteria for Web applications. ASP.NET includes several features designed to minimize both the number of failures and the impact of failures when they do occur. Frequent garbage collection (removal of unused code from memory) prevents performance degradation caused by insufficient memory. System management tools monitor performance and can automatically restart failing processes. ASP.NET minimizes the impact of failures with centralized session-state management that can hand off failing processes without loss of data or disruption to the user.

Debugging and Error Handling

Even the best programmers occasionally write programs with errors. Errors can also be caused by incorrect user inputs, hardware failures, network failures, and myriad other causes. A good programmer anticipates such problems and writes programs that respond gracefully when they occur. ASP.NET provides programmers with a number of powerful tools for debugging and error handling. ASP.NETs Try-Catch blocks allow programmers to write custom error-handling routines for problematic code, such as database accesses. When errors occur Try-Catch automatically redirects program execution to specific error-handling code blocks. The feature can be used to provide a graceful response to the user and to report the problem to the system administrator.

ASP.NET's trace object provides developers with a useful tool for debugging and fine-tuning code during development. The trace object provides detailed analysis of code execution, including order of execution, CPU time, and memory usage (Homer, 2000).

FUTURE OF ASP AND ASP.NET

ASP.NET is the cornerstone of Microsoft's .NET strategy and the company has committed tens of millions of dollars to promoting it. It is a full-featured enterprise-scale development framework that provides Web developers with a rich array of powerful tools for creating Web sites and Web services. It is still in its infancy but there is little doubt that it will play an important role in the future of Web development.

The introduction of ASP.NET makes the long-term outlook for ASP uncertain. The large base of existing applications virtually assures that Microsoft will continue to support ASP for many years. Although ASP does not support many of ASP.NET's advanced capabilities, it is an easy-to-use technology and an excellent developmental tool for Web sites that do not require ASP.NET's advanced features. Both technologies have made a significant contribution to the advancement of server-side programming technologies.

GLOSSARY

Browser Software programs that run on the user's computer that are used to view Web pages.

Caching (pronounced CASH-ing) A technique computers use to save frequently accessed files.

Client An application that runs on a computer that relies on a server to perform some operations.

CLR Common language run time; responsible for managing code within Microsoft's ASP.NET framework. It also manages security and other tasks.

IIS Internet information server by Microsoft Corporation, one of the most widely used Web server applications on the market. It incorporates all the tools required by high-traffic commercial Web sites, such as security, extensions, logging, and database interfaces. It is included with several versions of Microsoft's Windows operating system.

MSIL Microsoft intermediate language is an intermediate level of code compilation between source code and machine code. It facilitates ASP.NET's ability to support a wide number of programming languages and operating systems.

.NET A flexible platform introduced by Microsoft that allows programs written in different languages to be compiled to run under .NET environments. The .NET platform can be used to develop programs for the Web, for desktops, for PDAs, or for any other Web-enabled device. ASP.NET is part of .NET and runs on Web servers.

Script An executable list of commands created by a scripting language, such as VBScript, JavaScript, PHP, Perl, or JScript.

Server A computer that serves Web pages and other files to a client via the Internet.

Tag Intermixed with text to describe the document's structure or its visual formatting. They are used with markup languages such as HTML and XML.

View State A hidden control used by ASP.NET to store form data in an encrypted format between page loads.

Web Services Technologies that allow computer applications to communicate via the Internet. Web services are built upon XML and SOAP technologies.

XML Etensible markup language; allows users to create their own tags to describe data content.

CROSS REFERENCES

See *ActiveX; Client/Server Computing; Extensible Markup Language (XML); HTML/XHTML (HyperText Markup Language/Extensible HyperText Markup Language); Web Services.*

REFERENCES

Bidgoli, H. (1999). *Handbook of management information systems: A managerial perspective*. London: Academic Press.

Buckman, R., & Bulkeley, W. (2002). IBM, Microsoft battle again, this time over "Web services." *Wall Street Journal*, May 10, p. B1.

Gates, B. (June 18, 2001). *Letter to technology professionals: Why we're building .NET technology*. Retrieved April 25, 2002, from http://www.microsoft.com/presspass/misc/06–18BillGNet.asp

Homer A., Sussman, D., Anderson, R., & Howard, R. (2000). *A preview of Active Server Pages* +. Birmingham, UK: Wrox Press.

Kiely, D. (2000). *Microsoft solidifies its .NET plans*. Retrieved January 19, 2002, from http://www.informationweek.com/805/prmicrosoftnet.htm

Mitchell, S., & Atkinson, J. (2000). *Sams Teach Yourself Active Server Pages 3.0 in 21 days*. Indianapolis, IN: Sams Publishing.

Mono Project (2002). Retrieved February 13, 2003, from http://www.go-mono.com

Payne, C. (2002). *Sams Teach Yourself ASP.NET in 21 Days*. Indianapolis, IN: Sams Publishing.

Sholler, D. (2002). *.NET seen gaining steam in dev projects*. Retrieved April 29, 2002, from http://techupdate.zdnet.com/techupdate/stories/main/0,14179,2860227,00.html

Sussman, D., Homer, A., Howard, R., Watson, K., Francis, B., & Anderson, R. (2001). *Professional ASP.NET*. Birmingham, UK: Wrox Press.

Ullman, C., Cornes, O., Libre, J. T., & Go, G. (2001). *Beginning ASP.NET using VB.NET*. Birmingham, UK: Wrox Press.

ActiveX

Roman Erenshteyn, *Goldey-Beacom College*

INTRODUCTION

ActiveX is a well-established technology that Microsoft developed in the early 1990s, and the ActiveX technologies have become an essential part of Microsoft applications and tools. Briefly, ActiveX is a term used to refer to a wide range of client/server technologies and components. ActiveX controls, formerly known as OLE (Object Linking and Embedding) objects, are components that one can insert into a Web page or other application. ActiveX controls enhance Web design with sophisticated features, such as advanced formatting or animation.

ActiveX components are incorporated in Microsoft's Component Object Model, which allows objects to interact with each other. ActiveX controls are an integral part of all current Windows operating systems. The technology is popular, and, because ActiveX is language independent, many traditional development tools (but not all of them) can build and deploy ActiveX controls. One of the most important and attractive features of ActiveX is its reusability. ActiveX controls extend the concept of classes, which was introduced in object-oriented programming languages to enable a programmer to reuse code easily in later projects. ActiveX takes this concept further, allowing programmers to package a control, reuse it in later applications, distribute it as a solution to other developers, or even develop a composite control—a control over other controls.

Thousands of ActiveX controls can be found on the Internet; some of them are freeware and some shareware, so a developer can easily download and include them into an application or Web design. ActiveX controls run in the same memory space as a browser. They must be downloaded and registered on the user's computer before the controls can be used. A number of ActiveX controls come as built-in objects with Windows. Other sets of ActiveX controls come with Internet Explorer, Microsoft FrontPage, ActiveX Control Pad, Microsoft Visual Studio 6.0, Microsoft Visual Studio.NET, and so on.

WHAT IS AN ACTIVEX CONTROL?

Historical Notes

Chappell and Linthicum (1997) reviewed processes and reasons that led to the development of ActiveX. It started with the introduction to OLE. The first release of Object Linking and Embedding (OLE 1) was developed as a technology that enabled users to create and to work with compound documents. A compound document contains elements created by two or more applications. For example, a user can combine pictures created using Adobe PhotoShop, text created using Microsoft Word, and a spreadsheet created using Microsoft Excel under a single document created using Microsoft PowerPoint. It enables users to focus more on the information they have to work on and less on the applications they use. The idea behind this was to link several documents together or to embed one document in another. The second version, OLE 2, provided users with the solution to a more general problem: how to connect different software components, that is, how software of different types should interact. This generalization produced what was called Component Object Model (COM). COM "establishes a common paradigm for interaction among all sorts of software—libraries, applications, system software, and more" (Chappell, 1996). Then Microsoft revised the definition of OLE, and this term was applied to anything that has been built using COM.

In the beginning of 1996, the definitions were changed again. Microsoft chose a new name—ActiveX. Primarily, this new term was associated with technologies related to the Internet and its applications. ActiveX was directly connected to OLE. Microsoft redefined the term OLE to mean only the technology used for creating the compound documents. As a result, new COM-based technologies labeled by OLE are now tagged with ActiveX.

An introduction to ActiveX controls can be found at http://msdn.microsoft.com/library/default.asp?url=/workshop/components/activex/intro.asp. ActiveX is defined as simply another term for "OLE object," or, more specifically, "COM object." A control is a COM object that at least supports the IUnknown interface and allows any control to be as lightweight as possible. Controls should actually support an interface, as well as methods, properties, and events.

OLE Control

ActiveX controls are often defined as OLE controls that support additional features. Chappell (1996) specified that OLE refers only to the technologies for creating compound documents: for embedding documents created by one application within a document created by a different application. For example, an Excel spreadsheet can be embedded within a Word document. In this case, the word processor doesn't need to build in the functions of a spreadsheet. The standard interfaces defined by OLE enable this interaction among all sorts of applications. Applications act as containers or servers, so one application always serves as the container (Word, for example) and another (Excel) acts as a server, which can place its documents within the container's document. Using OLE, a server's document can be linked to or embedded in the container's document.

The minimal requirement for an OLE control is a support of IUnknown and IClassFactory interfaces; however, an OLE control also supports a number of other interfaces that provide additional features, such as automation (methods, events, and properties) or user interface for the control. Every interface inherits from IUnknown that contains only three methods, the most important of which is QueryInterface. The IClassFactory interface allows programmers to create instances of a particular class. The following list is a sample OLE interfaces:

- IOleObject—supports communication between container and the control
- IOleInPlaceObject—supports in-place activation
- IDataObject—supports data transfer
- IDispatch—supports a control's methods and properties
- IPersist—includes six interfaces that support functions that enable a control to read or write its persistent data to storage, stream, or file
- IExternalConnections—supports functions that a control uses to track external connections
- IPerPropertyBrowsing—supports functions that allow container to retrieve individual control properties
- The complete description of these and other OLE interfaces can be found in MSDN on-line library (see references)

Additional features supported for ActiveX controls are initialization security, scripting security, run-time licensing, reduced footprint (for quick downloading), and digital certification.

COM Control

There are many sources of technical information, resources, and training on COM and COM-based technologies available on the Web. The Microsoft Developer Network (MSDN) is the first place to look for information on all Microsoft developer resources and Internet technologies (MSDN Library, n.d.). Some conceptual presentations from the 1998 Microsoft Professional Developers Conference in Denver are available (Microsoft, 1998). The Component Object Model defines how objects interact with each other. Charlie Kindel from Microsoft (1998) defined ActiveX simply as "A marketing name for a set of technologies and services, all based on the Component Object Model." ActiveX controls are using COM technologies to provide interaction with other types of COM components and technologies. Microsoft COM is the most widely used component software model. It provides a variety of integrated services, easy-to-use tools, and a lot of applications. Maloney (1998) defined COM as "a framework for creating and using components with a wide choice of tools, languages and applications." COM provides choice in the area of security because it provides common interface (SSPI, security support provider interface) where security providers can be plugged in. COM also provides the major elements necessary for technology to succeed: a solid specification and a single reference implementation that has been ported to multiple platforms (Maloney, 1998).

ACTIVEX FAMILY

ActiveX Controls

One of the most attractive features of ActiveX technology is the variety of ways that is components can be created. Figure 1 illustrates the ActiveX family, including tools that can be used to develop controls:

Chappell (1996) defined an ActiveX control as "a software component that performs common tasks in standard ways." The ActiveX controls specification specifies a standard set of interfaces that COM objects can support to carry out particular actions. Components need to interact with the code that uses them, and the ActiveX

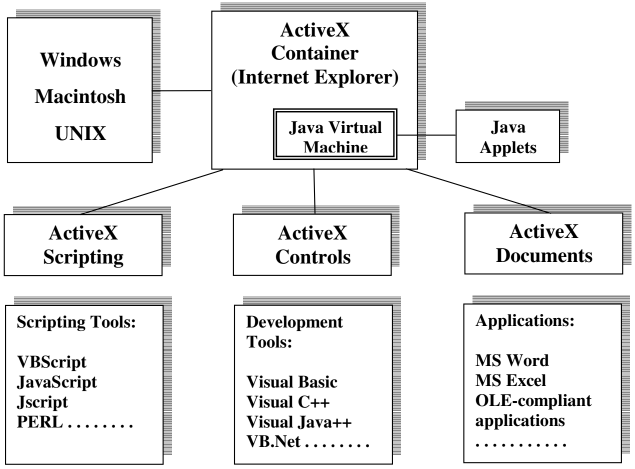
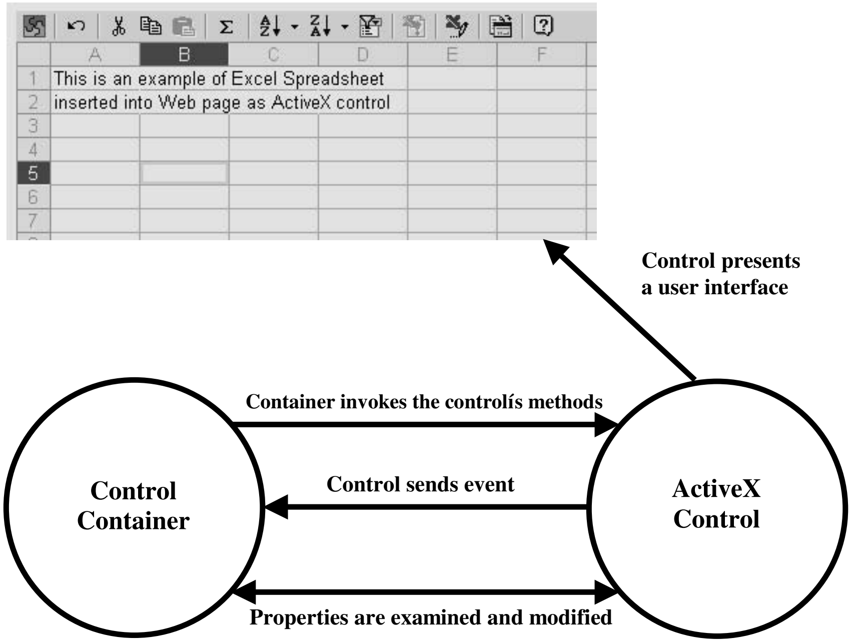


Figure 1: ActiveX family.



This is an example of Excel Spreadsheet
inserted into Web page as ActiveX control

A
B
C
D
E
F
1
2
3
4
5
6
7
Control presents a user interface
Container invokes the controlís methods
Control sends event
Control Container
ActiveX Control
Properties are examined and modified

Figure 2: ActiveX control's functionality.

controls specification and also determines rules for creating so-called control containers. A control container is client software that knows how to use an ActiveX control (Chappell, 1996). Applications developed using different computer languages, such as Visual Basic and C++, Java, can serve as control containers. Web browsers, such as Microsoft Internet Explorer, also are examples of containers. Netscape Navigator is not originally an ActiveX container; therefore, Web pages containing ActiveX controls may not function properly in Netscape browsers.

ActiveX controls for the Web are simple objects on a Web page that provide a user interface and support these interfaces. They add an interactive component to the page and allow users to communicate with the page. The code for a control can also be stored on a Web server and downloaded and executed when the browser or Web page requests them. An ActiveX control is compiled code with file extension.ocx.

Chappell (1996) also described the functionality of ActiveX control's specification as consisting of four main tasks:

1. Providing a user interface
2. Allowing the container to invoke control's methods
3. Exposing specific properties of the control and supporting modification of these properties
4. Examining the control's properties and modifying them

Each task is implemented by its own group of interfaces. This functionality is shown in Figure 2 (Chappell, 1996).

A variety of interfaces are implemented by an ActiveX control. This set of interfaces is available to ActiveX control developers so they can choose what they need. For example, Chappell (1996) described standard interfaces for an ActiveX control, and they support the control's functionality, which includes user's interface, methods, events, and properties. There are also a set of standard interfaces that ActiveX control containers support.

When the control is developed, it can be added to the Web page using the hypertext markup language (HTML) object tag `<OBJECT> </OBJECT>` and its attributes. (HTML is a simple language used to display information.) The following six attributes are important parts of control (Gilmore, 1999): CLASSID, CODEBASE, ID, WIDTH, HEIGHT, and BORDER. The CLASSID attribute contains the control's class ID, a unique identification tag generated when the control is compiled, and this allows recognizing and referencing the control. On a user's computer the control is then registered in the system registry with the class ID. The CLASSID attribute is the only one that is required; all others are optional. The following is an example of the `<OBJECT>` tag with its attributes for the Microsoft Animation Button Control. This button is an alternative to the common Command Button, and the Animated Button has a moving picture on it. In addition, the user can specify various start and end frames for conditions such as "mouse move," "mouse click," or "got focus" (the user presses the tab key until the button becomes the object on the document with the "focus"). This permits the user to display various visuals based on what the user is doing in relation to the button.

```
<OBJECT ID = "anbtn1" WIDTH = 127 HEIGHT = 71
       CLASSID = "CLSID:0482B100-739C-11CF-
         A3A9-00A0C9034920">
       <PARAM NAME = "_ExtentX" VALUE =
         "3360">
       <PARAM NAME = "_ExtentY" VALUE =
         "1879">
       <PARAM NAME = "defaultfrstart"
         VALUE = "0">
```

```
      <PARAM NAME = "defaultfrend" VALUE =
        "-1">
      <PARAM NAME = "mouseoverfrstart"
        VALUE = "0">
      <PARAM NAME = "mouseoverfrend"
        VALUE = "-1">
      <PARAM NAME = "focusfrstart" VALUE =
        "0">
      <PARAM NAME = "focusfrend" VALUE =
        "-1">
      <PARAM NAME = "downfrstart" VALUE =
        "0">
      <PARAM NAME = "downfrend" VALUE =
        "-1">
</OBJECT>
```

DefaultFrStart, DefaultFrEnd, DownFrStart, DownFrEnd, FocusFrStart, FocusFrEnd, MouseoverFrStart, and MouseoverFrEnd are the control's properties. The events for this control are Click, DblClick, Enter, Focus, Leave. Events are triggered when the user moves the mouse into the button (Enter) or off the button (Leave). Almost all editing tools today will automatically generate the correct HTML <OBJECT> tag code, including CLASSID information when ActiveX control is added to the Web page. The Class ID of a control can also be found by manually searching for the control in a system registry using the registry editor, REGEDIT.EXE. The CODEBASE attribute is used to specify a relative or absolute URL from where the control should be downloaded. An ID attribute works with the <OBJECT> tag the same as it works with other HTML tags, allowing the user to reference the element from scripting languages. The WIDTH, HEIGHT, and BORDER attributes specify the control's size and border. The <PARAM> tags, together with its NAME and VALUE attributes, are placed within <OBJECT> and </OBJECT> tags to assign initial values to a control's properties.

Gilmore (1999) also discussed the protection of controls from being used without permission. So-called run-time licenses allow controls to be executed, but not used for the development; design-time licenses allow controls to be used in the development stage. Different developers use different approaches to control licensing. Some developers require only design-time license, and some require both licenses. All the required licenses need to reside in the computer where the control is used. The License Manager ensures that run-time licenses (for controls that require them) are included in a license package file. The License Manager is included with Internet Explorer, and a reference to it should be included within the <OBJECT> tag. Another technology used for security is digital signature and code marking (Gilmore, 1999).

ActiveX Documents

ActiveX technology allows a user to work with embedded documents within a Web page as if the user were working in the original application. For example, a person using Web can activate an Excel spreadsheet, for example (see Figure 2), and use all Excel commands. ActiveX documents technology is based on OLE documents. The Web browser can serve as an ActiveX document container, whereas applications like Word, Excel and others are ActiveX document servers. Supporting ActiveX documents requires a number of additional user interfaces for an embedded document. For example, an interface is required to allow users to print the document in the same way it would be printed from the server application. Internet Explorer relies completely on ActiveX documents. For example, an HTML page is downloaded, the browser interprets it and displays the page, and, if a user downloads an Adobe Acrobat file, the browser also displays that information. A user may view the browser as one application, but the browser is actually built from several components that are linked together. Web browser objects can host any ActiveX document server, and it treats different servers (HTML viewer, Word, Excel, PowerPoint, Adobe Acrobat, etc.) identically.

ActiveX Scripting

Web browsers support HTML. Advanced user interface controls are not implemented. As the standard for creating a complete interface, Microsoft has provided ActiveX Scripting. Hillier (1996) defined ActiveX scripting as a language-independent standard that defines the relationship between a scripting host and a scripting engine. A scripting engine is an ActiveX component that conforms to the ActiveX specification (for example, Java Virtual Machine). A scripting host is an application that uses the scripting engine (an example of scripting host: Internet Explorer). The script is placed inside the actual HTML code between two tags <SCRIPT> and </SCRIPT>. After script is loaded into the engine, the host runs it.

Producing interactive Web pages is a primary goal for Web developers. Scripting languages is just one method of providing interactivity, but the use of scripting languages offers a number of advantages for developers. Scripts can be defined as a set of program instructions that a developer can embed in the HTML code, which is then executed by the browser. The most popular scripting languages are JavaScript and VBScript, but ActiveX Scripting is an open standard that allows other scripting languages to be used.

Server-Side ActiveX Controls

Dietel, Dietel, and Nieto (2001) described how server-side script or Active Server Pages (ASP) functionality can be extended with server-side ActiveX components. These types of ActiveX controls reside on the Web servers and lack a graphical user interface. Some of these controls included with Internet Information Server and Personal Web Server are the following:

- MSWC.BrowserType—ActiveX control for gathering information about the client's browser
- MSWC.AdRotator—ActiveX control for rotating advertisements on a Web page
- MSWC.NextLink—ActiveX control for linking Web pages together
- MSWC.ContentRotator—ActiveX control for rotating HTML content on a Web page

- MSWC.PageCounter—ActiveX control for storing the number of times a Web page has been requested
- MSWC.Counters—ActiveX control that provide general-purpose persistent counters
- MSWC.MyInfo—ActiveX control that provides information about a Web site Scripting.FileSystemObject—ActiveX control that provides an object library for accessing files on the server
- ActiveX Data Objects (ADO) Data Access Components—ActiveX controls that provide an object library for accessing databases

Java Virtual Machine

Java is a programming language that is similar to C++ but also provides exciting capabilities for the Internet and component technologies. Java Virtual Machine (Java VM) is implemented as an ActiveX control included with Internet Explorer and is used to execute Java programs and applets. Chappell (1996) specified that with Java VM these applets are treated as ActiveX controls, and because controls can be driven by scripts, Java applets can also be scripted. Then, using the ActiveX scripting interfaces, scripting languages can be used to access the methods exposed by an applet. In reality, when Internet Explorer encounters the `<APPLET>` tag, it internally converts it to an `<OBJECT>` tag with the CLSID of the Java VM's ActiveX control.

Security Features of ActiveX

As mentioned earlier, there are several security concerns related to ActiveX technology: initialization security, scripting security, run-time licensing, and digital certification.

After a control is initialized, it can receive data from a local or remote URL for initializing its state. Data can come from an untrusted source, and this is a security hazard. Several methods can be applied to ensure the control's security for initialization. Component Categories Manager creates the appropriate entries in the system registry. Microsoft Internet Explorer examines the registry before loading a control to determine whether these entries appear. Another method implements an interface IObjectSafety. If Internet Explorer determines that control supports this interface, it calls IObjectSafety::SetInterfaceSafetyOptions method before loading control to determine whether it is safe for initialization.

ActiveX controls can be accessed from scripts. Code signing can guarantee that the code is trusted. But even if a control is known to be safe, is not necessarily safe when automated by an untrusted script. The methods to ensure scripting security are identical to ones described earlier for initialization security.

Some ActiveX controls are distributed free of charge, but most of these should support design-time and run-time licensing. Design-time licensing ensures that a developer is building an application or Web page with a legally purchased control. In contrast, the run-time licensing ensures that a user is running an application or displaying a Web page that contains a legally purchased control. Design-time licensing is verified by control containers. Before these containers allow a control to be placed on Web page, they first verify that the control is licensed. This is achieved by calling certain functions in the control that validate the license. For run-time licensing the process is similar.

At the default security settings in Internet Explorer any object on an HTML page must be digitally signed. Digital signatures are created using Signing Code with Microsoft Authenicode Technology. It is a set of tools that can be downloaded from MSDN. A digital signature associates a software vendor's name and a unique public key with a file that contains an ActiveX object.

WHAT ARE THE ALTERNATIVES TO ACTIVEX?

Client-Side Scripting and Component Technologies

Component technologies are used in the Internet world to extend the functionality of Web pages. ActiveX is not the only way to accomplish this. Gilmore (1999) emphasized the role of component technologies in building e-business and other Web sites. Component technologies allow developers to create a flexible and user-friendly interface for Web pages. This interface can contain a large number of input fields, graphics, animations, event controls, and so on. As mentioned earlier, the component functionality is defined through the set of its properties, methods, and events. Components also make it possible to add a complex logic to a Web page in the same way it can be done for applications written in C++ or Visual Basic. The component technologies other than ActiveX control are Java applets and DHTML scriplets. Alternatives to component technology are client-side scripting and DHTML (Gilmore, 1999). Java applets can be created using Java language and they can run on any browser that has Java Virtual Machine, which serves as an interpreter for Java applet's code. DHTML scriplets (Gilmore, 1999) are Web pages with DHTML functionality. Similar to ActiveX controls, they can be used in COM applications, other then Web. DHTML scriplets are supported by Windows, Mac, and UNIX platforms.

ASP as a Server-side Alternative to ActiveX

There are several server-side alternatives to ActiveX, and most of these are covered in other chapters. Briefly, ASP (and its most recent version, ASP.NET; Kalata, 2002) is Microsoft's implementation of Web server programming technology. ASP makes it easier for Web developers to create dynamic Web applications. Active Server Pages are Web pages that contain server side scripting. When the client requests an ASP page, the Web server passes the request to the ASP application. The ASP application then detects if there are any global scripts that need to be processed. The ASP application inserts the code from any server-side include pages. ASP.NET is a language-independent technology that is used to develop Web applications. The two main types of Web resources created with ASP.NET applications are WebForms and Web Services. WebForms allow users to develop and process forms on Web pages, as well as to develop cross-browser Web applications. Web Services are ASP.NET pages that contain publicly exposed code so the other applications can interact with them. ASP applications run much faster than CGI scripts and can incorporate HTML pages and forms, scripts written in VBScript or JavaScript, as well as ActiveX components

ActiveX Strengths and Limitations

Gilmore (1999) provided a detailed analysis of the advantages and limitations of component technologies and client-side scripting, highlighting the main benefits of using components in the Web pages:

- **Robustness.** Traditionally used Web technologies cannot compete with what is possible using programming languages such as C++, Java, and Visual Basic. The code written in one of these languages can be included in a component and then into a Web page;
- **No duplicated scripts.** The same component can be used in different pages instead of duplicating scripts.
- **Maintenance and reusability.** In the case when several Web pages use the same component, its source code is stored only once. This modular approach makes maintenance simple. Code encapsulation in a single entity allows reusing the component in other applications.
- **Implementation hiding.** Developers need to know only what the component does, not how it does it. Because developers do not need to worry about the specifics of component's code, they only need to know how to interact with the component's interface.

The biggest limitation for ActiveX is that there many products on the market that lack of the tools necessary for designing ActiveX controls.

Chappell (1996) and Linthicum (1997) discussed the debate over ActiveX and Java applet component technologies. Gilmore (1999) offered a detailed comparison of ActiveX versus Java applets and versus DHTML scriplets and client-side scripting. ActiveX controls can be developed using different programming languages and are not limited to Web browsers; they can run on different platforms, including Macintosh and UNIX, as well. Java applets must be created using the Java language, but they can run on any platform that has a Java VM. This is possible because a Java VM shields the applets from the details of the host operating system. Today Java applets are used more frequently in Web applications than ActiveX controls. ActiveX controls have to register themselves with the Windows registry and execute as native Windows applications. This allows the controls access to all native features, such as file I/O (input/output) and devices and even memory.

ActiveX controls are fully supported by Internet Explorer and partially supported by Netscape Navigator. Users of Netscape Navigator need an ActiveX plug-in, and because most of them don't have it, they may prefer other alternatives. ActiveX controls run faster than alternative technologies because the controls are compiled into code. Client-side scripting and DHTML scriplets are the slowest options because they must be interpreted. The functionality of ActiveX is almost unlimited, while Java applets functionality is limited by Java language abilities. Client-side scripting and scriplets are also limited by scripting language features. Both ActiveX controls and Java applets provide similar levels of security. An important feature is the download speed. The download of ActiveX requires users to wait. ActiveX controls remain on the client, however, and therefore subsequent downloads are not required. The Java applet download is the slowest because class files have to be downloaded every time the Web page that contains an applet is opened. Client-side script download speed is the fastest. The set of skills required to create ActiveX controls includes C, C++, Visual Basic, Java, or other tools that can create them. The Java language is the only skill needed to develop a Java applet. All component technologies can be reused and maintained easily. ActiveX controls maintenance requires a development tool for the language used to create it. DHTML scriplets can be maintained with any text editor. An important feature is the security of business logic. The logic programmed using client-side script, DHTML, or DHTML scriplets can be seen using a Web browser. Source code developed for ActiveX control and Java applet is compiled and not accessible.

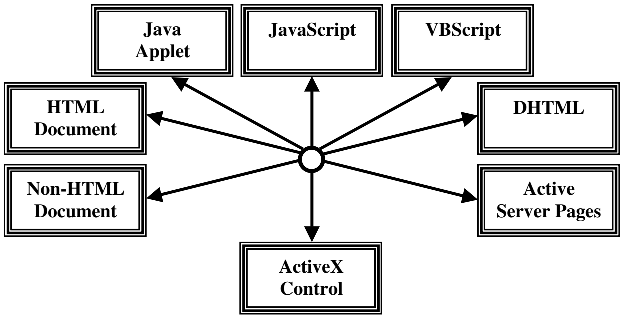
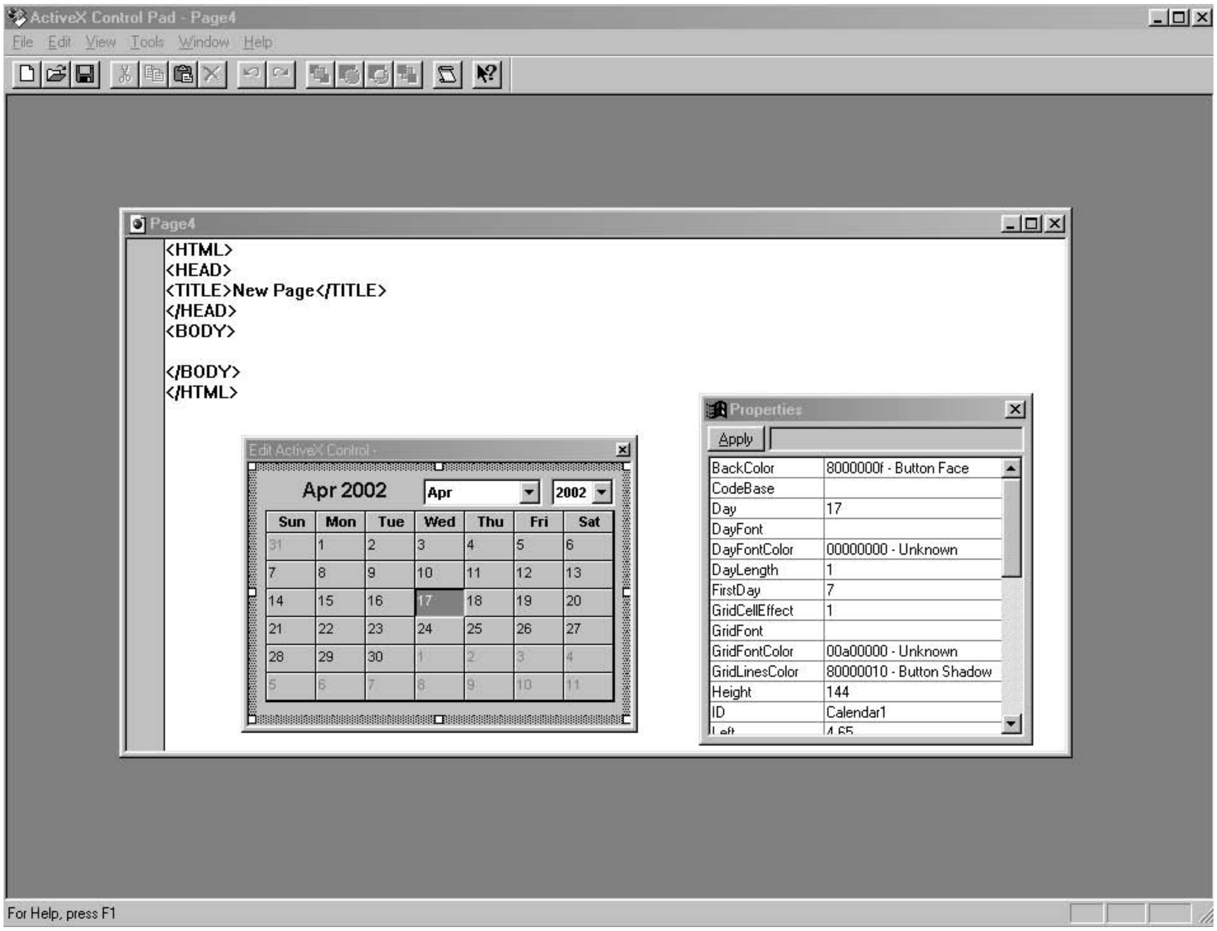


Figure 3: ActiveX integrates different Web technologies.

The answer to the question of which approach is better is simple. Developers must consider the strengths and weaknesses of ActiveX controls, Java applets, DHTML scriplets, and other technologies before making a choice. These technologies are not mutually exclusive; each is suited to particular situations. Nonetheless, one advantage ActiveX controls is obvious: They can be used in a wide variety of applications and are not limited to use in Web applications. This makes it easy to integrate and reuse a component from a variety of existing technologies used for Web design and development., a concept is illustrated in Figure 3.

ActiveX and Microsoft's .NET Framework

The next-generation platform introduced by Microsoft is called .NET, and it is closely related to COM and ActiveX. In general, .NET Framework "is a component of the Microsoft Windows operating system that provides the programming model for building, deploying, and running Web-based applications, smart client applications, and XML Web services" ("Top 10 Reasons," 2002). This technology has many features that improve existing component technologies, including COM and ActiveX. It promotes interaction with COM components and generates



```
<HTML>
<HEAD>
<TITLE>New Page</TITLE>
</HEAD>
<BODY>

</BODY>
</HTML>
```

Apr 2002 Apr 2002

Sun	Mon	Tue	Wed	Thu	Fri	Sat
31	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	1	2	3	4
5	6	7	8	9	10	11

BackColor	8000000f - Button Face
CodeBase	
Day	17
DayFont	
DayFontColor	00000000 - Unknown
DayLength	1
FirstDay	7
GridCellEffect	1
GridFont	
GridFontColor	00a00000 - Unknown
GridLinesColor	80000010 - Button Shadow
Height	144
ID	Calendar1
Left	4.65

Figure 4: ActiveX Control Pad.

a wrapper around existing components, so the programmer not needed to rewrite them. Applications built using the .NET Framework can connect with existing systems and packaged applications, regardless of their underlying platform, via XML Web services or via other system-specific connectors (“Top 10 Reasons,” 2002). The .NET Framework technology for interacting with data, Microsoft ADO.NET, is designed for today’s Web-based style of data access. ADO.NET is a new version of ActiveX Data Object Model that provides objects that allow programmer to interface with the database. With .NET Framework ADO.NET is installed, and XML is a standard that allows programmers to describe and store data within text files. ADO.NET also allows programmers to retrieve data from a variety of sources, including XML data. By using these technologies programmers can exchange data across products and platforms (Kalata, 2002). The future of .NET Framework is promising, but this is not directly related to the main subject of this chapter.

BASIC AND ADVANCED ACTIVEX TOOLS

Microsoft ActiveX Control Pad

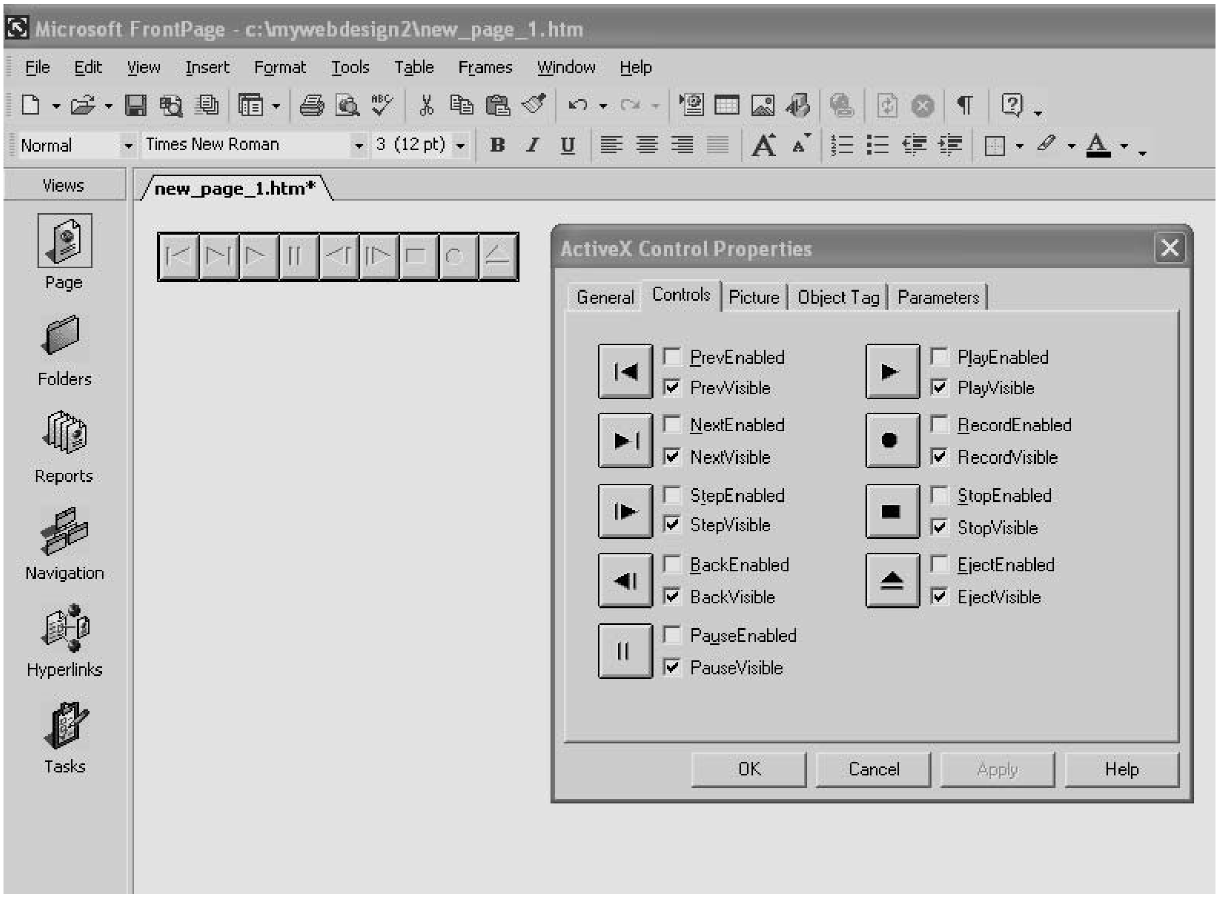
Microsoft ActiveX Control Pad is a smart Web development tool that allows adding ActiveX controls and ActiveX scripting (VBScript or Jscript) to Web pages. It also includes the Microsoft HTML Layout Control. Using this tool, a developer can include advanced layout and multimedia features, such as exact object placement, layering, and other effects. At the time of this writing, setuppad.exe file can be downloaded from the MSDN library page (msdn.microsoft.com/library/default.asp). It looks similar to HTML editors, extended to insert ActiveX controls. Figure 4 illustrates the use of ActiveX Control Pad to insert calendar control (shows the object and properties).

Microsoft FrontPage 2002

Microsoft Front Page also allows adding ActiveX controls to existing or new Web pages. To insert ActiveX control to the Web page, a developer has to open this page and then follow the following sequence: Insert—Web Component—Advanced Controls—ActiveX Controls. It opens a list of available controls. Figure 5 shows an example of inserting a control. The figure shows the Multimedia Control itself and a properties window that includes several tabs, such as General, Parameters, Controls (it is open), Object Tag.

Text and Graphics Editors

Important components of ActiveX Control development tools are decent text and graphics editors. The simplest



☐ PrevEnabled
☑ PrevVisible
☐ NextEnabled
☑ NextVisible
☐ StepEnabled
☑ StepVisible
☐ BackEnabled
☑ BackVisible
☐ PauseEnabled
☑ PauseVisible
☐ PlayEnabled
☑ PlayVisible
☐ RecordEnabled
☑ RecordVisible
☐ StopEnabled
☑ StopVisible
☐ EjectEnabled
☑ EjectVisible

Figure 5: Multimedia control inserted with Microsoft FrontPage.

text editor that can be used for this purpose is Notepad, which comes with Windows. HTML editors, such as CoffeeCup HTML Editor (www.cofeecup.com) or Web Edit, contain shortcuts for automatic insertion of a large number of HTML tags. There are also hundreds of graphics editors available on the market. The following list contains only a small sample of available graphics editors: Adobe PhotoShop, Microsoft Photo Editor, Microsoft Image Composer, ArcSoft Photo Studio, MGI Photo Suite, and Paint Shop Pro 6.2.

ActiveX Software Development Kit

The Microsoft Platform Software Development Kit (SDK) provides the documentation, samples, header files, libraries, and tools a user needs to develop Windows-based applications. The applications developed with this edition of the SDK can run on Windows .NET Server, Windows XP, Windows ME, Windows 2000, Windows 98, Windows 95, and Windows NT.

Microsoft Visual Basic Control Creation Edition

Visual Basic Control Creation Edition, a member of the Microsoft Visual Basic programming system family, is designed specifically to be the fastest and easiest way to create ActiveX controls. There are three general usage scenarios for creating ActiveX controls with the 1996 Control Creation Edition:

1. Creating ActiveX controls from scratch. Everything is included in the Control Creation Edition to allow the creation of complete, stand-alone ActiveX controls from scratch. It is expected that the following two usage scenarios will be more common, however.
2. Subclassing and customizing an existing ActiveX control. Developers can take advantage of the variety of commercially available and free ActiveX controls. An existing control can be subclassed, customized, and then compiled, creating a custom version of the same control.
3. Aggregating multiple ActiveX controls into a control "assembly." Developers can take advantage of the large market of commercially available ActiveX controls by aggregating multiple controls together into an ActiveX control project, customizing their look and behavior, and then compiling the group of controls together into a single control. The resultant control can then be inserted into a Web page or a client/server application to "wrap" or contain the entire user interface elements of that application.

Microsoft Visual Studio Visual Basic comes with a variety of ActiveX controls that can be easily included in different applications. The instructions and scenarios for almost all of them can be found in MSDN. The following is a list of several controls:

- Communications Control—provides an interface to a standard set of communication commands (establish a connection to a serial port, connect to another communication device, exchange data, monitor and respond to various events and errors)
- CoolBar Control—allows creation of user-configurable toolbars
- DayTimePicker Control—displays date and time information and acts as the interface through which users can modify this information (see Figure 4);
- DataRepeater Control—functions as a data-bound container of any other user control. It allows the user to create a catalog that includes images and descriptions of each product and to create bankbook applications to track personal finances (can be found in almost all banks' online services)
- ImageCombo Control—has the ability to include a picture with each item in the list portion of the combo
- Internet Transfer Control—implements hypertext transfer protocol (HTTP) and file transfer protocol (FTP), which allows a user to add an FTP browser to any application, to create an application that automatically downloads files from a public FTP site, to parse a Web site for graphics references and download graphics only, among other things
- MSChart Control—allows a user to plot data in charts according to specifications (e.g., dynamic data, such as current prices of selected stocks)
- Multimedia Controls—manages the recording and playback of MCI (Media Control Interface) devices (see Figure 5)
- PictureClip Control—creates an image resource bitmap that stores all the images needed for an animation

Microsoft Foundation Classes (MFC) and ActiveX Template Library (ATL)

ActiveX controls can be also developed using Visual C++. Microsoft offers some tools that help a C++ programmer to create an ActiveX control. MFC is a set of C++ classes that support COM, OLE, and ActiveX (among other things). An existing OLE control can be converted into ActiveX control adding additional features (typical for ActiveX), such as safety, run-time licensing, and digital certificates. This technology allows for creating small controls, but it requires the correct MFC dynamic link library to be installed.

An ActiveX Template Library (ATL) is a set of C++ template classes designed to create small and fast ActiveX controls (COM objects). If a control is developed using ATL, the end user needs to download only the object, because the control doesn't need a run-time dynamic link library to be resident. With ATL developers can create several control types. The use of ATLs minimizes the number of interfaces that a control needs; the control will draw itself. More detailed information about ATL and its downloads can be found at msdn.microsoft.com/visualc/prodinfo/.

Online ActiveX Services

Thousands of ActiveX Controls are available on the Web, including freeware that can be used for private purposes and is helpful for developers. The following are the links to the most attractive and useful sites at the time of this writing.

- **www.webdeveloper.com/activex** This site contains useful information and links for developers, including technical articles, downloads, development guides, and more.
- **download.cnet.com/downloads/0–10081.html** This site features ActiveX Controls grouped into the following clusters: application development, browser enhancements, control development, database connectivity, online applications, and tools and utilities.
- **browserwatch.internet.com/activex.html** This site is called BrowserWatch—ActiveX Arena. It is the place where users can find the links to all of the ActiveX Controls available on the Web. There are links to a variety of form design controls, multimedia controls, graphic controls, sound controls, document viewer controls, and productivity controls—to almost everything a user may need for Web design and development.
- **www.geocites.com/SiliconValley/Park/3545** This site features new and interesting ways to use ActiveX controls and contains directions to do-it-yourself, step-by-step guides on how to use these controls. The information and examples from this site are especially useful for beginners.
- **www.vision-factory.com/activex.htm** This site offers several ActiveX Controls downloads free of charge for private use, including rounded buttons control, text scrolling, Outlook-like user interface, a Windows Explorer–style tree, list view of files and directories, among others.
- **www.webexpressions.com/resource/registry/activex.cfm** This site contains helpful information about using ActiveX Controls on the Web as well as downloadable samples. Many of the samples are useful for almost any type of site; others of them are just for fun. Short explanations of how each one works and can be modified are included. There are also links to hundreds of sites that are using these to make their pages interactive. There are ActiveX Controls for simple and dual pop-up menus, cascading popup menus, text and button menus, marquees, stock and news tickers, info buttons and various text effects.
- **www.active-x.com** This is a large repository of ActiveX Controls developed by different companies. It includes a variety of useful controls (some are free). The following is the list of top downloads in mid-2002:
  1. TILISOFT—These Internet ActiveX Controls allow users to retrieve HTML pages from the Net, to post data to or retrieve data from an HTTP server, and to transfer files over the Net using Transmission Control Protocol and Internet Protocol.
  2. BetterButton—This replaces the standard Windows button to add many more effects, such as color, pictures, drop shadows, and various styles. It also creates

new styles that break from tradition, such as Visual Studio.NET and active (hover button) styles. It can create various picture styles, set alignment, wrap text, and more.

3. Active Image—These ActiveX Controls features include the ability to create or modify images on the fly, to perform text drawing with font styles and angles, and to use drawing tools such as rectangles, polygons, pie charts, arcs and circles, and linear image transformations.
4. Animation GIF ActiveX—This ActiveX Control allows users to display animation GIF files.
5. FreeLink ActiveX—This is a label control that is highlighted when the user moves the mouse over it. It will also start a URL in the user's default Web browser or e-mail client if it is clicked. It is freeware and includes complete source code.

- **www.vbfreeware.com/download.asp?type=ax66** This site offers ActiveX Controls for different tastes and purposes that can be downloaded for free. Controls are available for a variety of applications, including communication, multimedia, Web development, and more.
- **shop.store.yahoo.com/componentone-llc/comstudenful.html** ComponentOne Studio Enterprise is the most comprehensive collection of components for ActiveX, .NET, and ASP.NET. It is subscription service that provides more than 30 components, including grid components, reporting components, charting components, data components, user interface components, and e-commerce components.

ACTIVEX CONTROL EXAMPLES

The following examples have been taken from a Web site, Actives Controls examples (members.tripod.com/~activecontrol/index.htm, developed by Greg Jarol, University of British Columbia, Canada) and can be viewed there.

The structured graphics control comes as a default with Internet Explorer and allows the user to create powerful three-dimensional graphics using simple vector primitives that can then by animated using scripting (such as VBScript or JavaScript). This particular example (Figure 6) uses JavaScript.

Source code for structured graphics control:

```
<OBJECT ID = "piechart"
     CLASSID = "clsid:369303C2-D7AC-11d0-
       89D5-00A0C90833E6"
     STYLE = "position:relative; Width:125;
       height:100">
     <PARAM NAME = "Line0001" VALUE =
       "SetFillColor(0,0,255)">
     <PARAM NAME = "Line0002" VALUE =
       "SetFillStyle(1)">
     <PARAM NAME = "Line0003" VALUE =
       "SetFont('Verdana',40,650,0,0,0)">
     <PARAM NAME = "Line0004" VALUE =
       "Text'Arc',40,650,0,0,0)">
     <PARAM NAME = "Line0005" VALUE =
       "SetLineColor(0,0,0)">
     <PARAM NAME = "Line0006" VALUE =
       "SetLineStyle(1,3)">
     <PARAM NAME = "Line0007" VALUE =
       "SetFillColor(0,0,255)">
     <PARAM NAME = "Line0008" VALUE =
       "SetFillStyle(1)">
     <PARAM NAME = "Line0009" VALUE =
       "Pie(-150,-70, 200,200,15,22,0)">
</OBJECT>
<SCRIPT LANGUAGE = "JavaScript1.2">
     if (document.all)
            setInterval("piechart.
              Rotate(5,5,5)",100)
</SCRIPT>
```

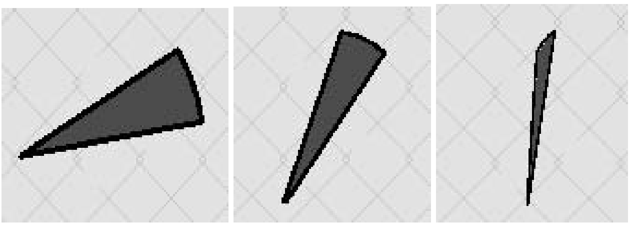
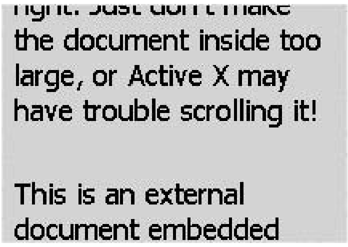


Figure 6: Example of the structured graphics control.



right. Just don't make the document inside too large, or Active X may have trouble scrolling it!

This is an external document embedded

Figure 7: Example of the document scroller.

The document scroller (Figure 7) control allows you not only to embed another HTML document inside the main document, but also to have the "internal" document scroll up-down or left-right all by itself. Note that there are no scroll bars.

Source code for document scroller:

```
<OBJECT ID = test2 CLASSID = "clsid:
  1A4DA620-6217-11CF-BE62-0080C72EDD2D"
     WIDTH = 220 HEIGHT = 150>
     <PARAM NAME = "szURL" VALUE =
       "page1.htm">
     <PARAM NAME = "ScrollDelay" VALUE = 7>
     <PARAM NAME = "ScrollPixelsY"
       VALUE = -2>
     <PARAM NAME = "ScrollPixelsX"
       VALUE = 0>
     <PARAM NAME = "LoopsY" VALUE = -1>
     <PARAM NAME = "DrawImmediately"
       VALUE = 1>
</OBJECT>
```

Figure 8: Screen capture example of the path control.

The path control (Figure 8) is an advanced control available in Internet Explorer that allows the user to animate visual objects (images, text, etc.). Objects can move around inside the document freely following complex, predetermined paths. The example in Figure 8 animated an HTML form button.

Source code for path control:

```
<INPUT TYPE = BUTTON NAME = btnSpline
  VALUE = "Flying button!"
      STYLE = "position:absolute;LEFT: 20;
        TOP: 200">
<OBJECT ID = "pthPolygon" CLASSID = "CLSID:
  D7A7D7C3-D47F-11D0-89D3-00A0C90833E6">
      <PARAM NAME = "AutoStart" VALUE = "-1">
      <PARAM NAME = "Repeat" VALUE = "-1">
      <PARAM NAME = "Bounce" VALUE = "0">
      <PARAM NAME = "Duration" VALUE = "10">
      <PARAM NAME = "Shape" VALUE =
        "Polygon(8,50,100,50,150,100,100,100,
        250,50,250,50,275,25,150,0,150)">
      <PARAM NAME = "Target" VALUE =
        "btnSpline">
</OBJECT>
```

The drop-down menu control (Figure 9) creates a contextual menu that can be further enhanced with extra functionality, such as going to the selected URL when clicked on. VB Script is used for this control.

Source code for the drop-down menu control:

```
<OBJECT ID = "IEMenu1" CLASSID = "clsid:
  7823A620-9DD9-11CF-A662-00AA00C066D2"
      WIDTH = 1 HEIGHT = 1 ALIGN =
        LEFT HSPACE = 0 VSPACE = 0>
      <PARAM NAME = "Menuitem[0]" VALUE =
        "This is the first item">
      <PARAM NAME = "Menuitem[1]" VALUE =
        "This is the second item">
      <PARAM NAME = "Menuitem[2]" VALUE =
        "This is the third item">
      <PARAM NAME = "Menuitem[3]" VALUE =
        "This is the fourth item">
```

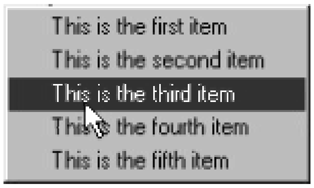


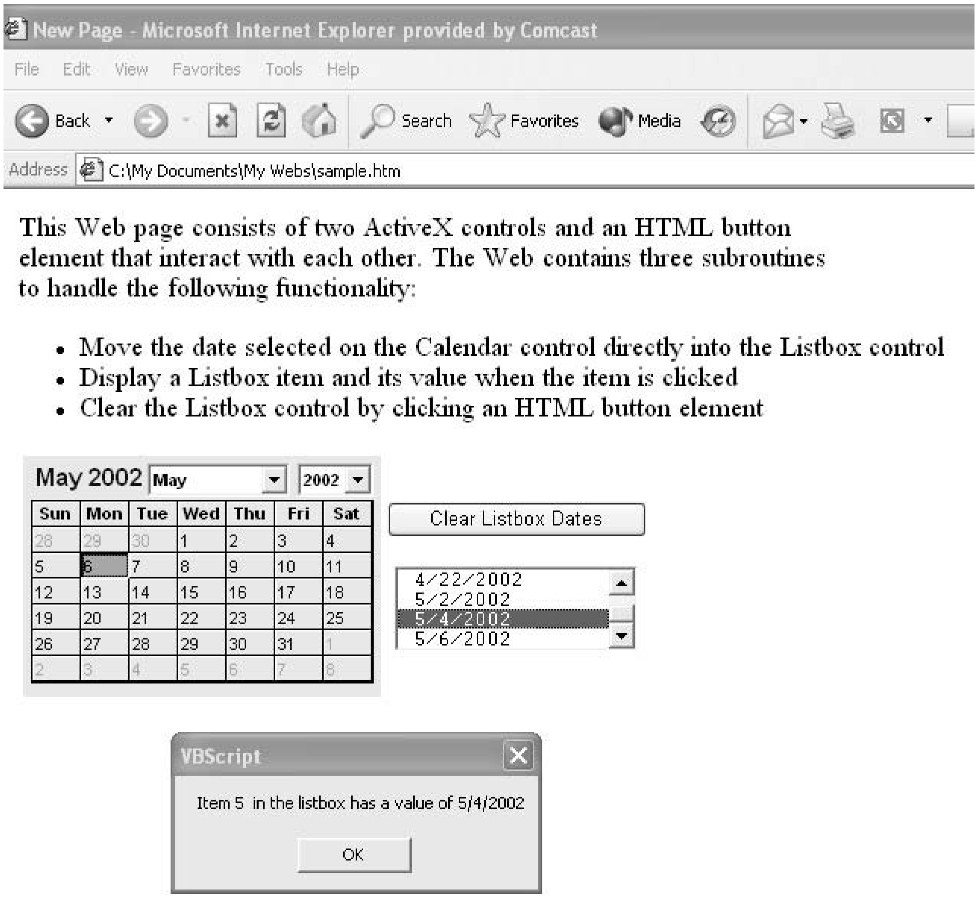
Figure 9: Screen capture example of the drop-down menu control.

```
      <PARAM NAME = "Menuitem[4]" VALUE =
        "This is the fifth item">
</OBJECT>
<SCRIPT LANGUAGE = "VBScript">
      sub IEMenu1_Click(ByVal x)
             Alert "You clicked on menu item:
               "& x
      end sub
      sub ShowMenu_onClick
              call IEMenu1.PopUp()
      end sub
</SCRIPT>
<DIV onClick = "IEMenu1.PopUp()">
  <B>Drop down menu!</B></DIV>
```

Gilmore (1999) offered another interesting and useful example of ActiveX Controls. Often Web page contains a number of ActiveX controls that interact with each other. In this example, a Web page contains two controls—a calendar control and a list box control, as well as an HTML button. The calendar and list box ActiveX controls are included into Microsoft Visual Studio. The Web page is shown in Figure 10.

When the date is checked on the calendar, it appears in the list box. This happens because the calendar's Value property is passed as a parameter to the list box's AdItem method, and this takes place in an event handler for the calendar's Click event. The VBscript for is as follows:

```
sub Calendar1_Click()
       ListBox1.AddItem Calendar1.Value
end sub
       The list box entries can be cleared
by clicking on an HTML button, the OnClick
event of the button invokes the ClearList
subroutine:
sub ClearList()
        ListBox1.Clear
end sub
        Finally, a list box item and its
value are displayed in a message box when
that item is clicked:
sub ListBox1_Click()
        msgbox "Item" & ListBox1.ListIndex
          + 1 & " in the list box has a
          value of " &
               ListBox1.Value
end sub
        The control for the ListBox1:
<OBJECT ID = "ListBox1" WIDTH = 160
  HEIGHT = 60
          CLASSID = "CLSID:8BD21D20-EC42-
            11CE-9E0D-00AA006002F3">
          <PARAM NAME = "ScrollBars" VALUE =
            "3">
          <PARAM NAME = "DisplayStyle"
            VALUE = "2">
          <PARAM NAME = "Size" VALUE =
            "4233;1588">
          <PARAM NAME = "MatchEntry" VALUE =
            "0">
```


New Page - Microsoft Internet Explorer provided by Comcast
File Edit View Favorites Tools Help
Back
Search
Favorites
Media
Address
C:\My Documents\My Webs\sample.htm
Clear Listbox Dates
VBScript
OK

This Web page consists of two ActiveX controls and an HTML button element that interact with each other. The Web contains three subroutines to handle the following functionality:

- Move the date selected on the Calendar control directly into the Listbox control
- Display a Listbox item and its value when the item is clicked
- Clear the Listbox control by clicking an HTML button element

May 2002 May 2002

Sun	Mon	Tue	Wed	Thu	Fri	Sat
28	29	30	1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	1
2	3	4	5	6	7	8

4/22/2002
5/2/2002
5/4/2002
5/6/2002

Item 5 in the listbox has a value of 5/4/2002

Figure 10: Web page with multiple ActiveX controls (Gilmore, 1999).

```
        <PARAM NAME = "FontName" VALUE =
          "Courier">
        <PARAM NAME = "FontHeight" VALUE =
          "200">
        <PARAM NAME = "FontCharSet" VALUE =
          "0">
        <PARAM NAME = "FontPitchAndFamily"
          VALUE = "2">
        <PARAM NAME = "FontWeight" VALUE =
          "0">
</OBJECT>
```

This example clearly shows how simple it is to develop multiple controls on a Web page and how they can communicate with each other.

CONCLUSION

ActiveX controls have grown from OLE technology and become very efficient for use in a number of different containers. A variety of software development tools, as well as user-productivity tools facilitate the coding of ActiveX controls. ActiveX controls make it easy for software developers to create, reuse, and integrate different software components. ActiveX allows Web developers to build multimedia rich and productive Web sites quickly. Web browsers serve as ActiveX containers, and therefore Web page developers are able to include ActiveX objects and documents. These objects establish a variety of useful interfaces that make a Web page interactive. This is especially important for e-commerce Web sites on which interaction with customers is a key feature. ActiveX technology is flexible; it can easily be combined with other Web design technologies such as HTML, Java applets, JavaScript, VBScript, and others. There are thousands of ActiveX controls available online as freeware or shareware. Web development software and tools, such as Microsoft FrontPage or Macromedia Dreamweaver, make it easy to include ActiveX controls in Web pages.

GLOSSARY

ActiveX A broad range of client/server technologies that are part of Microsoft's Component Object Model. ActiveX technology is frequently incorporated on Web pages but can also be used in a variety of applications in the Windows environment (and occasionally in the Mac environment but not in Unix or Linux environment).

ActiveX Control A component that can be easily inserted into a Web page to enhance design. Controls can also be used in other Microsoft's applications not related to Web.

ActiveX Documents The document contained in an ActiveX container (Internet Explorer). Whereas a traditional document (such as a Word document) is static, ActiveX documents need not be. Using Visual Basic, the user can create a complete application with the semantics of a traditional document. In other words, users have the functionality of the application but the flexibility of a document. Thus, the "document" is truly active. The concept can be found at http://msdn.microsoft.com/library/default.asp?url=/library/en-us/vbcon98/html/vbcondocumentobjectsthefutureofforms.asp

ActiveX Scripting Any use of VBScript and Jscript within Internet Explorer that controls the integrated behavior of active controls.

Class A data structure that specifies the properties and behavior of objects.

Component Object Model (COM) Microsoft's framework for developing and supporting program component objects.

Component A unit of compiled or scripted code that encapsulates a set of functionalities (e.g., a set of functions).

Container An application that defines the outermost document that embeds another document; an application that can run components.

Event A notification sent by one object to another that an action has occurred. The receiving object executes an associated method.

Java Virtual Machine A component that enables a Web page to run Java applets.

Method Code that is executed when an object receives an event message.

Object An instance of a class. The object consists of the data variables (properties) declared in the class definition along with any methods or procedures that act on this data.

Object Linking and Embedding (OLE) Microsoft's framework for a compound document technology. Briefly, a compound document is something like a display desktop that can contain visual and information objects of all kinds: text, calendars, animations, sound, motion video, continually updated news, controls, and so forth. Part of Microsoft's ActiveX technologies, OLE takes advantage and is part of a larger, more general concept, the Component Object Model.

OLE custom control (OLX) A special-purpose program that can be created for use by applications running on Microsoft Windows systems. The file extension for ActiveX controls is .olx.

Properties An attribute (or characteristic) of a control (size, fonts, color, etc.).

CROSS REFERENCES

See *Active Server Pages; Client/Server Computing; Java; Visual Basic; Visual Basic Scripting Edition (VBScript).*

REFERENCES

Chappell, D. (1996). Understanding ActiveX and OLE. Redmond, WA: Microsoft Press.

Chappell, D., & Linthicum, D. S. (1997). It's invasive. It's ubiquitous. But what, exactly, is ActiveX? *Byte, 9*. Retrieved November 5, 2002 from www.byte.com

Deitel, H. M., & Deitel, P. J., Nieto, T. R. (2001). *e-Business & e-Commerce. How to program.* Upper Saddle River, NJ: Prentice Hall.

Gilmore, S. (1999). Extending the HTML User Interface with Components: In the Microsoft Commerce Solutions. *Web Technology.* Redmond, WA: Microsoft Press.

Hillier, S. (1996). Inside Microsoft Visual Basic, scripting edition. Redmond, WA: Microsoft Press. (Also available at MSDN Library; see URL at MSDN Library [n.d.]).

Kalata, K. (2002). Introduction to ASP.NET. Boston, MA: Thomson—Course Technology.

Kindel, C. (1998). *ActiveX and the Web architecture. Technical overview.* Presented at the Professional Developers Conference 98, Denver, Colorado. Retrieved November 5, 2002, from www.microsoft.com/com/presentations/default.asp

Linthicum, D. S. (1997). Java and ActiveX. *Byte, 9*. Retrieved November 5, 2002 from www.byte.com

Maloney, J. (1998). "COM+" Building on the Success of the Component Object Model. Presentation at the Professional Developers Conference, Denver, Colorado. Retrieved from November 5, 2002, from www.microsoft.com/com/presentations/default.asp

Microsoft Corporation (1998). Microsoft COM. Component Object Model. Presentation at the Professional Developers Conference 98 in Denver, Retrieved November 5, 2002, from www.microsoft.com/com/presentations/default.asp

MSDN Library (n.d.-a). ActiveX controls. Retrieved November 18, 2002, from msdn.microsoft.com/library/default.asp and http://msdn.microsoft.com/library/default.asp?url=/workshop/components/activex/intro.asp

Top 10 reasons for developers to use the .NET framework. (2002, July 23). Retrieved November 5, 2002, from msdn.microsoft.com/netframework/productinfo/topten/default.asp

FURTHER READING

ActiveX Controls. Retrieved November 18, 2002, from www.microsoft.com/com/tech/activeX.asp

ActiveX Control examples. Retrieved November 18, 2002, from members.tripod.com/~activecontrol/index.htm

ActiveX unofficial guide. Retrieved November 18, 2002, from www.shorrock.u-net.com/activex.html

Box, D. (1998). *Essential COM.* Reading, MA: Addison-Wesley.

Box, D. (1998, March). Q&A ActiveX/COM. *Microsoft Systems Journal*. Available online at http://www.microsoft.com/msj/defaultframe.asp?page=/msj/0398/activex0398.htm

Campise, F. J. (1999). COM primer. In *Microsoft commerce solutions*. Web Technology. Redmond, WA: Microsoft Press.

D. Chappell Associates Web Site. Retrieved November 18, 2002, from www.chappellassoc.com

Kirtland, M. (1998). Interface and component design with COM. Presented at the Professional Developers Conference 98, Denver, Colorado. Retrieved November 5, 2002, from www.microsoft.com/com/presentations/default.asp

Mansfield, R. (1997). The comprehensive guide to VBScript: The encyclopedic reference to VBScript, HTML & ActiveX. Retrieved November 18, 2002, from www.space154k.stai.com/library/vbscript/subjects/activex.htm

Parihar, M., Ahmed, E., Chandler, J., Hatfield, B., Lassen, R., McIntyre, P., Wanta, D. (2002). *ASP.NET bible*. New York: Hungry Minds Inc.

Rogerson, D. (1996). *Inside COM*. Redmond, WA: Microsoft Press.

What is an ActiveX document? Retrieved November 18, 2002, from http://msdn.microsoft.com/library/default.asp?url = /library/en-us/vbcon98/html/vbcondocumentobjectsthefutureofforms.asp

ActiveX Data Objects (ADO)

Bhushan Kapoor, *California State University, Fullerton*

INTRODUCTION

The field of Web applications development has seen significant changes over the past several years. Early Web applications consisted of simplistic static Web pages with text, graphics, and hyperlinks to other Web pages and had no database access. Over time, Web applications were developed that were dynamic in nature and could access traditional databases. Many organizations also developed e-commerce applications to do business on the Internet. These applications utilized databases to perform such tasks as storing and tracking customer information, purchases, and preferences, as well as monitoring company inventory and updating product catalogs and sales tax rates. Customers are now able to remotely access e-commerce applications and databases at negligible communication costs. Overall, these applications and databases have brought tremendous benefit and cost savings to the organizations that have installed them.

Having realized the gains from these new Web-based applications, organizations want to expand their Web applications base. They want to enrich and expand their information systems not only by converting or rewriting their existing traditional client/server applications to make them Web enabled, but also by creating new Web applications to take advantage of new opportunities created by intranets and the Internet. Organizations want to convert most, if not all, of their existing applications to Web-based applications because they are keenly aware that it will be very productive if their employees and business associates, including customers, suppliers, and salespeople, could access company applications and databases and do their work remotely as efficiently and cheaply as they could locally. In addition, they want to write additional Web-based applications that exploit the opportunities offered by their intranets and the Internet.

This is the information age, and the size and complexity of information continues to grow. Further, organizations have their important information distributed in various forms and locations. The information is more than likely stored not just in different types of traditional databases, such as Access, MS SQL Server, DB2, Oracle, and Sybase, but also in other data stores, including e-mail files, legacy flat files, spreadsheets, and Web-based text and graphics files. The latter set of data stores is likely to contain unstructured information, and new strategies and technologies are needed to access the information they contain. The easy and seamless exchange of information, irrespective of its form and data store, both within an individual organization and between various organizations, increases efficiency and effectiveness. This chapter intends to explain Microsoft's strategy for universal data access (UDA) and the technologies, including ActiveX data objects (ADO), used for implementing this strategy.

Over the past several years, Microsoft has developed or patronized several tools and technologies to access data from a wide range of data stores in order to comply with its UDA strategy. These technologies range from an early technology, namely open database connectivity (ODBC), to the Object Linking & Embedding Database (OLE DB), to the latest ActiveX data objects. Following a general description of these Microsoft technologies, this chapter will present a detailed discussion of ActiveX data objects. Microsoft UDA technologies can be effectively applied to both traditional client/server applications and Web-based applications. In this chapter, however, we concentrate primarily on ADO and its role in developing Web-based applications.

This chapter is divided into several sections. Following the introduction, we introduce the three generations of UDA technologies. In the next section, we take up the first generation of UDA technologies, including ODBC. Next, we take up the second-generation UDA technology, namely OLE-DB. The next four sections are devoted to the third-generation UDA technology, namely ADO. We devote next section to a brief discussion of advantages of the ADO technology. Following that, we discuss the basic elements of the ADO object model. We also take up, in that section, built-in ADO objects and collections, and their significant properties and methods. Next, we discuss the role of ADO in developing Web-based applications written with active server pages (ASP). The following section focuses on combining HTML, VB Script, ASP, and ADO for developing interactive Web-based applications. This topic is explained with the help of several examples. The final section summarizes the topics covered in this chapter.

UNIVERSAL DATA ACCESS

UDA is a strategy employed by Microsoft that is designed to provide a comprehensive means of accessing data from a wide range of data stores across intranets or the Internet. The data stores may be traditional structured databases, such as relational databases, or nonstructured data stores, such as text documents, spreadsheets, images, or pictures, supplied by various vendors. These data stores may also reside on heterogeneous computing platforms, including Windows-based and non-Windows-based operating systems. In summary, UDA provides high-performance, easy-to-use, transparent, and seamless access from a variety of client devices, programming languages, and scripts to a wide range of data stores, irrespective of their type, location, or platform.

Unlike other approaches, this strategy is designed to access data directly from their sources. Alternate approaches to UDA, called datacentric database strategies, require that data be transformed or converted into a common format, location, and platform before can be accessed by applications. UDA applications do not require costly data transformations or conversions and, therefore, accelerate development.

Because UDA is based on open industry specifications, it enjoys broad industry support. The UDA technologies can be grouped into three generations. The first generation of UDA technologies include open database connectivity, remote data objects (RDO), and data access objects (DAO). The second generation UDA technology consists of the Object Linking & Embedding Database. The third, and latest, generation of UDA technologies is composed of ActiveX data objects.

Although the first-generation technologies were developed before Microsoft formally announced its UDA strategy, they contain several important features of the UDA approach and they continue to be widely used in the industry. The second-generation technology, OLE DB, significantly extends the functionality of the first-generation technologies. Similarly, the third, and most recent, generation of UDA technologies, ADO, significantly extends the functionality present in the first two generation of UDA technologies.

ADO is built upon and works in close conjunction with some first- and second-generation technologies, especially ODBC and OLE DB. Because ODBC, OLE DB, and ADO are the primary and interrelated UDA technologies, Microsoft has bundled them together to a unified software package called the Microsoft data access components (MDAC). The MDAC software package may be downloaded from http://www.microsoft.com/data/default.htm.

FIRST-GENERATION UDA TECHNOLOGIES—ODBC, RDO, AND DAO

The ODBC technology continues to be a very important and successful data access standard. It provides a common interface to data stored in almost any relational database management system (DBMS) or even some flat-file systems, including ISAM/VSAM file systems. ODBC uses structured query language (SQL) as a standard means of accessing data. This has enabled applications, through a common set of codes built into the SQL language, to access information stored in any database or file that has an ODBC driver. Such a data store is sometimes called a SQL database, an ODBC data source, or simply a data source. Use of the SQL language and ODBC drivers has enabled applications to be independent of the data sources. Thus, a developer can build and distribute a client/server or a Web-based application without targeting a specific data source.

The following are the key software components of ODBC.

ODBC API: A library of function calls, error codes, and SQL syntax for accessing data.

ODBC database drivers: A set of dynamic link library (DLL) programs that process ODBC API function calls for specific data sources.

ODBC driver manager (ODBC32.DLL): Loads ODBC driver(s) specific to the data source(s) and keeps track of which database drivers are connected to which data sources. When an application creates a connection to a data source, this component works in the background, completely transparent to the application.

ODBC cursor library (ODBCCR32.DLL): Resides between the ODBC driver manager and the ODBC database drivers and handles cursoring within recordsets. We discuss cursoring and recordsets in detail later.

ODBC administrator: Allows the configuring of a database management system or a flat-file system to make it available as a data source for an application. Microsoft has included this important tool within its Windows operating systems. We discuss it in further detail later.

An application achieves independence from data stores by working through an ODBC driver written specifically for the database management system or the file system rather than by working directly with the system. The driver translates the ODBC API function calls and SQL into commands its database or file system can execute, making it available for a wide range of data stores.

Despite such advantages, ODBC has shortcomings. Two significant examples are the fact that ODBC is limited to relational database management systems and some flat-file systems, and ODBC API function calls are complex and difficult to use.

In response to the complexity of ODBC API function calls, Microsoft has developed two high-level programming models: data access objects and remote data objects. These high-level models have simplified the ODBC model and enhanced programmers' productivity. DAO and RDO, like ODBC, are limited to working with relational database management systems and flat-file systems only. But unlike ODBC, DAO and RDO are simple and easy to use.

DAO is designed to work primarily with file-server-based systems, such as Microsoft access database and legacy file systems. In a file-server-based system, the client (or workstation) is responsible for executing all processing logic components, including the presentation, the business, and the data access logic. The file-server's role is simply to provide shared access of files to the client. When a client requests particular data in a file, the server responds by sending an entire copy of the file. File-server models work well where file transfers are infrequent and file sizes are small.

Although DAO is designed to work with the file-server-based systems, RDO is developed to work with modern (client/server) database-server systems, such as MS SQL Server, DB2, Sybase, and Oracle databases. In a database-server model, the data access logic is executed at the database server and the presentation logic is performed at a client's computer.

In two-tier client/server applications, both the client and the database server participate in the execution of business logic execution. A portion of the business logic is executed at the server, while the remaining part is completed at the client.

In multi-tier applications, the client, database, and application servers are responsible for the execution of the business logic.

For Web-based applications, the business logic processing is shared among the Web server, database server, and client. The client-side processing code is executed at the client and the server-side processing code is executed at the Web server. The database server executes business logic using stored procedures. A stored procedure is made up of precompiled SQL statements that carry out a series of tasks on the database server. These procedures are written, stored, and executed on the database server. Although the file-server systems are mostly suited to small-scale applications, the database-server systems are designed for mission critical applications, requiring scalability, flexibility, maintainability, and better performance.

Although DAO is optimized to work with file-server-based systems, later versions of DAO also support client/server database systems. However, the performance is not satisfactory. RDO, on the other hand, is tuned for database-server systems and is well suited for developing large-scale multi-tier applications.

SECOND-GENERATION UDA TECHNOLOGY—OLE DB

OLE DB is a low-level programming interface to diverse sources of data, ranging from flat files, to relational databases, to object-oriented databases. OLE DB is based on Microsoft's Object Linking & Embedding and on the component object model (COM), both of which provide applications with uniform access to diverse data sources. Another important advantage of the OLE DB model is its high-performance design and support for multi-tier client/server and Web-based applications.

There are three key software components of OLE DB: OLE DB data providers, OLE DB services, and OLE DB data consumers.

OLE DB data providers are software components that allow access to diverse data stores. These include structured data stores, such as relational databases, and unstructured data stores, such as e-mail files, legacy flat files, spreadsheets, and Web-based text and graphics files. They provide access to each data store within a standard level of uniformity and functionality.

The OLE DB programming interface interacts with OLE DB data providers similarly to the way the ODBC API interacts with ODBC drivers—both OLE DB data providers and ODBC drivers provide direct access to specific data stores. Although there are a large number of ODBC drivers available for most popular database management systems, OLE DB data providers are available for a limited set of popular data stores, such as Microsoft access, MS SQL server, Oracle, and Exchange server. However, Microsoft has written an OLE DB provider for ODBC drivers, effectively providing OLE DB programmers access to both ODBC and OLE DB data source. Thus, OLE DB applications can be written for data stores that have ODBC drivers but no OLE DB providers.

OLE DB services are components that extend the functionality of data providers. Some database systems, for example, may lack cursoring features in their native data provider software. An OLE DB service, called the Microsoft cursor service, will add cursor functionality, allowing OLE DB data consumer applications to use this feature, irrespective of whether the native data provider software supports this feature.

OLE DB data consumers are software components that consume OLE DB data. Important data consumers are high-level data access models, such as ActiveX data objects.

THIRD-GENERATION UDA TECHNOLOGY—ADO

OLE DB, like ODBC, is a low-level, complex interface for accessing data. Although ODBC provides access to any SQL data store, OLE DB provides universal data access to SQL and non-SQL data stores. Just as DAO and RDO were developed to simplify access to SQL data stores, ADO was developed to provide a simple and easy-to-use interface for SQL and non-SQL data stores. ADO was developed as an OLE DB data consumer and provides a consistent high-level interface between applications and data stores. A traditional client/server or a Web-based application interacts with ADO to provide data required by the application.

The are several key advantages to ADO.

Programming language independence: ADO can be used with a wide range of languages, including Visual Basic, C++, Delphi, Java, J++, Java Script, Jscript, and VB Script. It can be used with any language that supports OLE and COM.

Support for traditional and Web-based applications: For developing traditional client/server and Web-based applications, ADO is an excellent mechanism for accessing data. It is required for developing Web-based applications with scripting languages, such as Java Script, Jscript, and VB Script, whenever they need to access data.

Easy to use and learn: ADO is easy to use and learn because it is based on a COM object interface, and it provides the ability to perform data queries and data manipulation using familiar commands, such as select, insert, update, and delete.

Universal data: ADO applications can access data from any OLE DB and ODBC data source. Applications that use ADO can use the OLE DB interface without losing any of the underlying OLE DB functionality.

ADO OBJECT MODEL

The ADO object model is made up of five top-level built-in objects: Connection, Recordset, Command, Record, and Stream. In addition to these top-level objects, ADO also contains four subordinate collections: Parameters, Fields, Properties, and Errors. These ADO collections contain Field, Property, Parameter, and Error objects, respectively. A collection consists of one or more objects referred to collectively.

Each top-level ADO object can exist independent of the other top-level objects. These top-level objects may contain subordinate collections that, in turn, contain their associated objects. For example, the Connection object has two subordinate collections, Errors and Properties, which, in turn, contain Error and Property objects. Similarly, the Recordset object has two subordinate collections, Fields and Properties, which, in turn, contain Field and Property objects (see Figure 1).

Each ADO object has two types of properties: ADO-defined built-in properties and provider-defined dynamic properties. ADO-defined built-in properties are available to the ASP (active server pages) programmer regardless of the particular data provider in use. Properties that are specific to a particular data provider are contained in the Properties collection as Property objects. Each top-level object has its own Properties collection.

In the ADO model, the subordinate objects relative to their parent object cannot be created independently. The top-level object must be created before its subordinate objects are created. The only exception is the Parameter object. The Parameters collection may be created independently of the Command object; however, it must be associated with a Command object before it can be utilized. A brief description of each ADO top-level objects and some of its associated collections and objects follows below.

Connection Object

The main purpose of the Connection object is to establish a link between an application and a data source. Once a connection is established the Execute method of this object can also be used to execute SQL queries, commands, and stored procedures against the target data source. For each query, a Recordset object is created and the result of the query execution, called a recordset or a cursor, is stored in this object. Recordsets, or cursors, are further examined later in this section.

The Connection object is also responsible for retrieving errors. An Error object contains details about a single operation involving data access. The Error collection contains all the Error objects.

In establishing a link between an application and a data source, the Connection object uses one of the following two types connections: a DSN connection or a DSN-less connection.

DSN Connection

A DSN connection is established in two steps: (a) Create an ODBC data source name (DSN), (b) execute "Open" method of the connection object. The DSN connection is created outside the ADO application using the ODBC administrator program. The ODBC administrator program is referred to by different names in different Windows operating systems. The program is called ODBC Data Sources (32-bit) on a Windows 95 or 98 systems, ODBC on a Windows NT system, and Data Sources (ODBC) on a Windows 2000 system. The shortcut used to execute the ODBC administrator program is located in the control panel (or administrative tools folder within the control panel) of the Windows operating system. To set up an ODBC DSN, the ODBC administrator program requires the following information: a unique name that represents the DSN; the selection of an ODBC driver, such as the

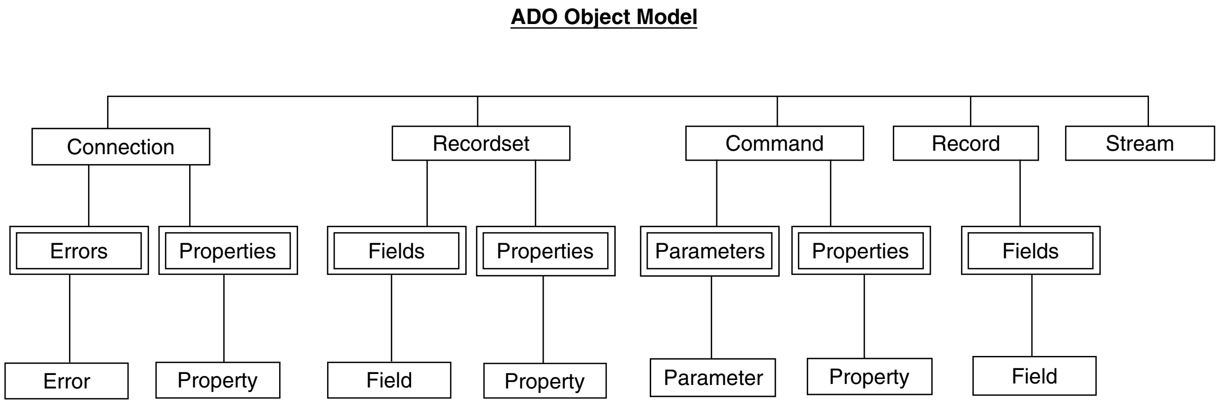


Figure 1: ADO object model.

Microsoft access driver (*.mdb), Microsoft ODBC for Oracle, SQL server, or Microsoft Excel driver (*.xls); he name of the data store (database or data file) to which the DSN will connect; and additional information, such as UserId and password, if required by the data provider to access the data store.

The Selection of a DSN Type. There are three types of DSNs: user DSN, system DSN, and file DSN. User DSN is limited to one user. The only user who will be able to connect to the data store with a user DSN is the user who creates it. A data store with a system DSN is available to all authorized users of the machine, including Windows NT and Windows 2000 services. A data store with a file DSN is available to each user who has the same driver installed. User and system DSN information is stored in the Window's registry. File DSN information is not part of the Window's registry but, instead, is held as a separate text file.

After creating an ODBC DSN, the ASP programmer may establish a DSN connection by passing the name of the ODBC DSN as a parameter to the "Open" method of the Connection object.

DSN-less Connection

In some cases the programmer may not be able or authorized to get physical or remote access to execute the ODBC administrator program needed to set up a DSN connection. This is especially true in the case of Web applications development. Developers of Web applications, generally, do not have physical or remote access for establishing a DSN connection on the Web server or the database server. In such cases, a DSN-less connection is suitable.

A DSN-less connection can be established explicitly or implicitly. An explicit DSN-less connection is established by providing connection information through the ConnectionString property of the Connection object. This information would vary depending on the data store and whether the connection uses an ODBC driver or an OLE DB provider, but it is similar to the information provided when setting up DSN connections in the ODBC administrator program. An implicit DSN-less connection is established with the Recordset object or the Command object.

Some of the methods built into the Connection object include Open, Close, and Execute. The Open method establishes a connection to the target data source while the Close method terminates the link. The Execute method allows the ASP programmer to perform basic SQL database operations, such as select, insert, delete, and update, on the data source specified by the Connection object. The Connection object is also capable of executing simple stored procedures. For stored procedures with dynamic parameters, the Command object is more suitable.

Some of the built-in properties of the Connection object are CommandTimeout, ConnectionString, ConnectionTimeout, Provider, State, and Version. The CommandTimeout property stores a number representing the number of seconds the Connection object will wait for a response from the data source after using the Execute method. The default CommandTimeout value is 30 seconds. If the Connection object does not receive a response within the CommandTimeout interval, it terminates the command and generates an error. The ConnectionTimeout property is similar to the CommandTimeout property but contains the time interval, in seconds, that the Connection object will wait for a connection after using the Open method. The default ConnectionTimeout value is 15 s. If no connection is established during the ConnectionTimeout period, the Connection object terminates the command and generates an error. The Provider property holds the name of the OLE DB provider for the Connection object. The State property is a read-only property that indicates whether the connection is open and the Version property is a read-only property that represents the ADO version number.

Recordset Object

Recordsets or cursors allow one to navigate through records and change data. Forward and backward navigation through records is performed in relation to the current record in a recordset. By default, the first record in a recordset is its current record. A recordset can be created with the help of the Recordset, Connection, or Command object. A recordset is made available when the Open method of the Recordset object is executed, the Execute method for the Command object is executed, or the Execute method for the Connection object is executed.

In the first two methods, there may be either an implicit or an explicit connection to the data store. In the third method, an explicit connection to the data store must be set prior to creating the recordset. In general, when using multiple recordsets within the same program, greater efficiency is achieved by creating an explicit connection to the data store prior to creating recordsets. Unlike implicit connections, when a connection is explicitly created, you have a handle for closing the connection. The implicit connection continues to remain open until the connection times out.

The Recordset object also provides control over the type of cursor and locking mechanism used. There are four different types of cursors, or recordsets. The cursor type must be specified before creating a recordset. Each cursor type is assigned a unique number between 0 and 3, as described below:

Cursor Type	Description
0	Forward only or read only cursor
1	Keyset cursor
2	Dynamic cursor
3	Static cursor

Type 0 (the forward only or read only) cursor is the default cursor. You can step through the recordset sequentially one by one using the MoveNext method. This cursor type does not support scrolling or updating.

Type 1 (keyset) cursor is a set of keys only. The other information is not duplicated in the recordset. In this type, records updated or deleted by other users are reflected in the recordset. However, new additions are not reflected in the recordset.

Type 2 (dynamic) is a scrollable cursor. All changes such as additions, updates, and deletions made by other users are reflected in the recordset.

Type 3 (static) is also a scrollable cursor; however, changes made by other users are not reflected in the recordset.

There are also several different lock types that can be used on a recordset when making changes such as additions, updates, and deletions to its records. The lock type must be specified before creating a recordset. Each lock type is assigned a unique number between 1 and 4, as described below:

Lock Type	Description
1	Read only
2	Pessimistic
3	Optimistic
4	Batch optimistic

Type 1 (read only) is the default lock type. It returns a read-only recordset and its records are available to other users.

Type 2 (pessimistic) lock type returns an updateable recordset. Before a particular record is updated, it is locked to other users. Update consists of two operations: read and write. The record is locked before the read operation is performed and is released just after it is written back.

Type 3 (optimistic) lock type returns an updateable recordset. As stated earlier, update consists of two operations: read and write. The record is not locked until the read operation is completed and is released just after the write operation is completed.

Type 4 (batch optimistic) is similar to lock type 3 cursors but updates are made to a batch of records at a time.

Built-in methods of the Recordset object include AddNew, Clone, Close, Delete, Move, MoveFirst, MoveLast, MoveNext, MovePrevious, Open, and Update. AddNew creates a new record to be added to the recordset. Clone creates a new recordset identical to the current recordset. Close method closes the recordset. Delete method removes the current record. MoveFirst method moves the position of the current record to the first record. MoveLast method moves the position of the current record to the last record. MoveNext method moves the position of the current record to the next record. MovePrevious method moves the position of the current record to the previous record. Open method opens a recordset; this method populates a recordset with records. Update saves any changes made to the recordset.

Built-in properties of the Recordset object include BOF, CursorType, EOF, LockType, and MaxRecords. BOF returns a Boolean value of true if the location of the current record is moved in front of the first record of the recordset. Otherwise, BOF returns false. CursorType indicates the type of cursor used in the recordset, such as forward only or read only, keyset, dynamic, and static. EOF returns a Boolean value of true if the location of the current record is moved past the last record of the recordset. Otherwise, BOF returns false. LockType indicates the type of lock used in the recordset, such as read only, pessimistic, optimistic, and batch optimistic. MaxRecords control the maximum number of records to return to a recordset from a query.

The Fields collection is created when the Recordset object's Open method executes. Field objects contain the metadata regarding the columns in the Recordset or the Record object, such as the name, type, length, precision, and data values.

Command Object

The Execute method of the Command object, similar to the Execute method of the Connection object, allows one to issue SQL queries, commands, and stored procedures against the target data source. However, the Command object is most useful when there is need to use a parameterized command or a stored procedure in an application repeatedly, passing in different values for the parameters each time. A collection of Parameter objects exposes the parameters associated with a Command object, based on a parameterized command or a stored procedure.

Record Object

The Record object is essentially a one-row Recordset object. The Record object can be created by (a) the execution of a query, a command, or a stored procedure that returns one row of data, (b) a row obtained from a Recordset, or (c) a row returned directly from a provider. Record objects facilitate access to information stored in nontraditional data stores. The row returned from the nontraditional data store may contain unstructured data, such as e-mail files, Web-based text, and directories, subdirectories, and files in a file system or folder. The directories, subdirectories, and files represented by Record objects can be copied or moved to another location with its CopyRecord or MoveRecord methods. Similarly, the directories, subdirectories, and files represented by Record objects can be deleted with its DeleteRecord method.

Stream Object

The Stream object provides the means of reading, writing, and managing a stream of bytes or text. The Stream object can be created from (a) a URL pointing to a file containing a stream of bytes or text data, (b) the default Stream object associated with a Record object, or (c) the independently created Stream object. The independently created Stream object is opened in memory. Data stream of bytes and text can be written to it and later saved in another stream or data store.

A stream of bytes or text data can be written to a Stream object with the Write and WriteText methods. Similarly, a stream of bytes or text data can be read to a Stream object with the Read and ReadText methods. The directories, subdirectories, and files represented by Record objects can be copied or moved to another location by its CopyRecord or MoveRecord methods. Similarly, the directories subdirectories, and files represented by Record objects can be deleted by its DeleteRecord method.

ADO AND ACTIVE SERVER PAGES (ASP)

The ADO provides a data model for Web-based applications that use active server pages (ASP). ASP is an open, compile-free, scripting model based on Microsoft technology. It is incorporated into and run as a part of

Microsoft Internet information server (IIS), which in turn is built into and runs on Windows NT server, Windows 2000 professional, or Windows 2000 server. ASP 3.0 is the latest version compatible with IIS 5.0 under Windows 2000 operating systems. Some earlier versions of ASP also run on the personal Web server (PWS), which in turn runs on Windows NT Workstation 4.0 or Windows 95/98 operating systems.

ASP allows client-side and server-side scripting in the same ASP page. It also contains text, HTML tags, and Java applets. However, only server-side scripts are executed on the server. The client-side code consists of scripts, ActiveX controls, and Java applets. These are sent to the client for its execution there. Two commonly used scripting languages are VB script and Java script. The default language of a server-side script is VB script. ASP interacts with data stores through ADO commands coded in server-side scripts.

The ASP files are saved as ASCII or text-only files and have an .asp file name extension. When the Web server receives a request for a Web page, the server first examines the file extension of the requested document. If the file extension is .asp, the Web server directs this ASP document to the ASP script host (ASP.DLL). The ASP script host sends out any server-side script it finds in the ASP file to the appropriate script engine. For example, the ASP script host sends server-side VB scripts to the VB script engine and server-side Java scripts to the Java script engine.

Each script engine interprets scripts submitted to it. When a script includes ADO commands, the script engine executes the command. The script engine returns an output consisting of HTML tags to the script host. The script host places all these outputs together and returns it, along with the client-side processing code, to the browser as an HTML stream.

ADO AND WEB-BASED APPLICATIONS DEVELOPMENT

Use the following steps to access and manipulate a database from an ASP Web page.

Create and open an ADO connection to a data store.

Create and open an ADO recordset from the data store.

Create one or more recordsets and manipulate their records.

Close the connection and recordsets.

These steps will be explained with the help of some examples. In these examples, the following information is used.

Database:	C:\group4\SampleDb.mdb
DSN name:	dsnDb
Driver:	Microsoft Access Driver (*.mdb)
Provider:	Microsoft.Jet.OLEDB.4.0
Table:	Customers
Fields:	CustName, CustPhone
User Id:	bkapoor
Password:	SECRET

Throughout these examples, <% and %> tags are used to indicate server-side scripts. The scripts included in these examples are written in VB script.

Step 1: Create and Open an ADO Connection

This step contains three examples, using DSN connection to a database, DSN-less connection with an ODBC driver, and DSN-less connection with an OLE DB driver.

DSN Connection to a Database

Before finalizing code that creates and opens a data source name (DSN) connection to the database, create a system DSN. A system DSN is created by the execution of the ODBC administrator shortcut from the control panel of the Windows operating system. During execution of the ODBC administrator program, the programmer will be asked to assign a unique name that represents the DSN. This name is referred to when creating a DSN connection. The following snippet contains code to create and open a DSN connection:

```
<%
Dim conn
Set conn = server.CreateObject
  ("ADODB.connection")
Conn.open "dsn=dsnDb; uid=bkapoor;
  pwd=SECRET;"
%>
```

DSN-less Connection to a Database

In this method, the programmer creates a connection string that contains information needed to create a connection to the database. The information includes an ODBC driver or an OLE DB provider, the database name, and user ID and password. The DSN-less connection to a database can be created either by using an ODBC driver or an OLE DB provider.

The following snippet contains code to create and open a DSN-less connection that uses an ODBC driver:

```
<%
Dim Conn, ConnString
Set Conn = Server.CreateObject
  ("ADODB.Connection")
ConnString = "DRIVER={Microsoft Access
  Driver (*.mdb)}; DBQ=C:\group4\
  SampleDb.mdb; "
Conn.ConnectionString=ConnString
Conn.Open
%>
```

The following snippet contains code to create and open a DSN-less connection that uses an OLE DB provider:

```
<%
Dim Conn, ConnString
Set Conn = Server.CreateObject
  ("ADODB.Connection")
ConnString="Provider=Microsoft.Jet.
  OLEDB.4.0; Data Source=C:\group4\
  SampleDb.mdb;"
```

```
Conn.ConnectionString=ConnString
Conn.Open
%>
```

Step 2: Create and Open an ADO Recordset to a Database

The following snippet contains code to create and open a DSN-less connection with an OLE DB provider, and a recordset:

```
<%
Dim Conn, ConnString, rs
set Conn=Server.CreateObject
  ("ADODB.Connection")
ConnString="Provider=Microsoft.Jet.
  OLEDB.4.0; Data Source=C:\group4\
  SampleDb. mdb;"
Conn.ConnectionString=ConnString
Conn.Open
set rs = Server.CreateObject
  ("ADODB.recordset")
dim CursorType, LockType
CursorType = 0
LockType = 1
rs.Open "Select * from Customers",
  Conn, cursorType, lockType
%>
```

Step 3: Create One or More Recordsets and Manipulate Their Records

This step contains five examples: read and display records from a recordset, read records from a recordset and display them as an HTML table, insert a record into a database, update a database record, and delete a database record.

The following ASP program contains code to read and display records from a recordset:

```
<HTML>
<HEAD>
  <TITLE>ADO EXAMPLE</TITLE>
</HEAD>
<BODY>
```

Here are results from the processed ASP page!

```
<%
Dim Conn, ConnString, rs
set Conn=Server.CreateObject
  ("ADODB.Connection")
ConnString="Provider=Microsoft.Jet.
  OLEDB.4.0; Data Source=C:\group4\
  SampleDb.mdb;"
Conn.ConnectionString=ConnString
Conn.Open
set rs = Server.CreateObject
  ("ADODB.recordset")
rs.Open "Select * from Customers", Conn
rs.MoveFirst
do until rs.EOF
  Response.Write rs("CustName") & " "
  Response.Write rs("CustPhone") & "<br>"
  rs.MoveNext
loop
rs.close
Conn.close
set rs=nothing
set Conn=nothing
%>
</BODY>
</HTML>
```

The following ASP program reads records from a recordset and displays them as an HTML table:

```
<HTML>
<HEAD>
<TITLE>ADO EXAMPLE</TITLE>
</HEAD>
<BODY>
```

The following are results from the processed ASP script:

```
<%
Dim Conn, ConnString, rs
set conn=Server.CreateObject
  ("ADODB.Connection")
ConnString="Provider=Microsoft.Jet.
  OLEDB.4.0; Data Source=C:\group4\
  SampleDb.mdb;"
Conn.ConnectionString=ConnString
Conn.Open
set rs=Server.CreateObject
  ("ADODB.recordset")
rs.Open "Select * from Customers", conn
rs.MoveFirst
do until rs.EOF
  Response.Write "<TABLE BORDER = '1'
    WIDTH ='100%'>"
  Response.Write "<TR><TD>" & rs("CustName")
    & "</TD>"
  Response.Write "<TD>" & rs("CustPhone")&
    "</TD></TR>"
  rs.MoveNext
loop
rs.close
Conn.close
set rs=nothing
set Conn=nothing
%>
</table>
</BODY>
</HTML>
```

The following ASP program contains code to insert a record into a database. After insertion, this program will display all its records as an HTML table.

```
<HTML>
<HEAD>
  <TITLE>ADO EXAMPLE</TITLE>
</HEAD>
<BODY>
```

The following are results from the processed ASP script:

```
<%
Dim Conn, ConnString, rs, rs1
set conn=Server.CreateObject
  ("ADODB.Connection")
ConnString="Provider=Microsoft.Jet.
  OLEDB.4.0;
Data Source=C:\group4\SampleDb.mdb;"
Conn.ConnectionString=ConnString
Conn.Open
SQLcmd = "INSERT INTO Customers(CustName,
  CustPhone) VALUES ('Bhushan Kapoor',
  '(714)333—3333')"
Set rs = Server.CreateObject
  ("ADODB.Recordset")
Dim cursorType1, lockType1
CursorType1 = 3
LockType1 = 3
rs.Open SQLcmd, conn, CursorType1,
  LockType1
set rs1 = Server.CreateObject
  ("ADODB.recordset")
rs1.Open "Select * from Customers",
  conn
rs1.MoveFirst
do until rs1.EOF
  Response.Write "<TABLE BORDER =
    '1' WIDTH ='100%'>"
  Response.Write "<TR><TD>" & rs1("CustName")
    & "</TD>"
  Response.Write "<TD>" & rs1("CustPhone")
    & "</TD></TR>"
  rs1.MoveNext
loop
rs.close
rs1.close
Conn.close
set rs=nothing
set rs1=nothing
set Conn=nothing
%>
</table>
</BODY>
</HTML>
```

The following snippet contains code to update a database record:

```
<%
SQLcmd = "UPDATE Customers SET CustPhone=
  "(310)222-2222"
WHERE CustName = 'John Smith'"
Set rs = Server.CreateObject
  ("ADODB.Recordset")
rs.Open SQLcmd, Conn, 3, 3
%>
```

The following snippet contains code to delete a database record:

```
<%
SQLcmd = "DELETE * FROM Customers WHERE
  CustName = 'John Smith'"
Set rs = Server.CreateObject
  ("ADODB.Recordset")
rs.Open SQLcmd, Conn, 3, 3
%>
```

Step 4: Close the Connection and Recordsets

In the final step, the connection and all its recordsets should be closed. Recordsets should be closed before closing the connection. After closing, one should also release the computer memory that was used to store these objects by setting their values to nothing. This process is sometimes termed garbage collection. The following snippet contains code to close the connection and its recordsets and also to release memory. In this snippet, the variable “Conn” has been used to represent a connection and two variables, rs and rs1, to represent two recordsets.

```
<%
rs.close
rs1.close
Conn.close
set rs=nothing
set rs1=nothing
set Conn=nothing
%>
```

SUMMARY

Progressive organizations have two major goals for their information systems. The first is to convert their existing applications to Web-based applications. The second goal is to write additional Web-based applications to take advantage of the opportunities created by intranet and the Internet.

This is an information age and the size and complexity of information continues to grow. Organizations have their important information distributed in various forms and locations. Universal data access (UDA) is a Microsoft strategy designed to provide a comprehensive means to access data from a wide range of data stores distributed across intranets or the Internet.

Microsoft has developed or patronized several technologies tied to its UDA strategy. These technologies can be grouped into three generations. The first generation of UDA technologies contains ODBC, RDO, and DAO. The second-generation UDA technology consists of OLE DB. The third generation is the latest generation of UDA technologies and contains ADO.

The ODBC provides a common interface for accessing data stored in almost any relational DBMS or even some flat-file systems. ODBC uses SQL as a standard language for accessing data. Microsoft has created two high-level programming models, DAO and RDO, to simplify the ODBC model. DAO is written to work primarily with file-server-based systems and RDO is designed for database-server systems. RDO is well suited for developing large-scale multi-tier applications.

OLE DB is a low-level programming interface to diverse data stores, including flat files, relational databases, and object-oriented databases. OLE DB provides applications with uniform access to diverse data sources. Another important advantage of the OLE DB model is its high-performance design and support for multi-tier client/server and Web-based applications.

The ADO object model is made up of five top-level objects, Connection, Recordset, Command, Record, and Stream; four subordinate collections, Parameters, Fields, Properties, and Errors; and four associated objects, Parameter, Field, Property, and Error, within its own collections. Each top-level object can exist independently of the other top-level objects. The subordinate objects/collections cannot be created independently of their parent objects.

The following steps may be used to access and manipulate a database from an ASP Web page: (a) create and open an ADO connection to a data store, (b) create and open an ADO recordset to a data store, (c) create one or more recordsets and manipulate their records, and (d) close the connection and recordsets.

ADO enjoys broad industry support because it provides a consistent, easy-to-use, high-level interface between applications and diverse data stores and does it for both traditional client/server and Web-based applications.

GLOSSARY

ADO (ActiveX Data Objects) ADO is the latest Microsoft data-access programming model. It provides a consistent high-level interface with many different types of data stores, including relational databases, flat files, e-mail files, spreadsheets, text documents, and graphics files.

ADO collections A collection consists of one or more objects referred to collectively. ADO model contains four collections: Parameters, Fields, Properties, and Errors. These collections contain Parameter, Field, Property, and Error objects, respectively.

ADO Object Model ADO object model is made up of five top-level objects (Connection, Recordset, Command, Record, and Stream), four subordinate collections (Parameters, Fields, Properties, and Errors), and four associated objects (Parameter, Field, Property, and Error) each within its own collection. Each top-level object can exist independently of the other top-level objects. The subordinate objects/collections cannot be created independently of their parent objects.

Cursor type There are four types of cursors: forward only or read only, keyset, dynamic, and static cursors. A cursor type must be set before opening a recordset.

DAO (Data Access Objects) DAO is a high-level programming model developed to simplify the ODBC programming model. DAO is designed to work primarily with file-server-based systems.

Database server system In a database server model for a Web based system, the data logic is processed on the database server, the presentation logic is processed on the client, and the business logic processing is shared between the Web server, database server, and client.

DSN connection DSN (data source name) connection between a Web server and a database server is established in two steps: (a) create an ODBC DSN and (b) execute the "Open" method of the connection object.

DSN-less connection There are two types of DSN-less connections: explicit and implicit. An explicit DSN-less connection is established through the ConnectionString property of the Connection object. An implicit DSN-less connection is established based on the Recordset object or the Command object.

File-server system In a file-server system, the client is responsible for executing data logic, presentation logic, and business logic. The file server's role is simply to provide shared access of data files to the client.

Lock type There are four cursor lock types: read only, pessimistic, optimistic, and batch optimistic. A cursor lock type must be set before opening a recordset.

MDAC (Microsoft Data Access Components) Microsoft provides MDAC in order to make universal data access possible. MDAC consists of three important technologies: ODBC, OLE DB, and ADO.

ODBC (Open Database Connectivity) ODBC technology provides a common interface to access data stored in almost any relational database management system and some flat-file systems, including ISAM/VSAM file systems.

OLE DB data consumers OLE DB data consumers are software components that consume OLE DB data. Important data consumers are high-level data access models, such as ADO.

OLE DB data providers OLE DB data providers are software components that allow one to access diverse data stores, including both relational and non-relational databases, with a standard level of uniformity and functionality.

OLE DB services OLE DB services are software components that extend the functionality of OLE DB data providers.

RDO (Remote Data Objects) RDO is a high-level programming model developed to simplify the ODBC programming model. RDO is written to work primarily with database-server-based multi-tier systems. RDO facilitates access to data stored in almost any SQL database.

Recordset A recordset or a cursor is a set of records that are the result of a SQL query, a command, or a stored procedure.

Stored procedure A stored procedure is made up of precompiled SQL statements that carry out a series of tasks. Stored procedures are stored and executed on the database server and are created to execute the data logic and some business logic.

UDA (universal data access) The UDA approach is a Microsoft strategy designed to provide a comprehensive means of accessing data from a wide range of data stores across intranets or the Internet.

CROSS REFERENCES

See *Active Server Pages; Client/Server Computing; Databases on the Web; Electronic Commerce and Electronic Business; HTML/XHTML (HyperText Markup Language/*

Extensible HyperText Markup Language); JavaScript; Structured Query Language (SQL); Visual Basic; Visual Basic Scripting Edition (VBScript).

FURTHER READING

ASP 101. Retrieved February 5, 2003, from http://www.asp101.com

Brinkster. Retrieved February 5, 2003, from http://www.brinkster.com

Crawford, C., & Caison, X., Jr. (1999). *Professional ADO RDS programming with ASP.* Wrox Press Ltd.

Deitel, X., Deitel, X., & Nieto, X. (2001). *E-business and E-commerce.* New York: Prentice Hall.

Goldman, J. E., Rawles, P. T., & Mariga, J. R. (1999). *Client/server information systems.* New York: John Wiley & Sons.

Gottleber, T. T., & Trainor, T. N. (2000). *HTML with an introduction to JavaScript.* New York: Inwin McGraw–Hill.

Gunderloy, M. (1999). *Visual Basic developer's guide to ADO.* San Francisco: Sybex.

Holzner, S. (1999). *ADO Programming in visual basic 6.* New York: Prentice Hall PTR.

JavaScript Source. Retrieved February 5, 2003, from http://javascript.internet.com

Kalata, K. (2001). *Internet programming (with VBScript and JavaScript).* Boston: Course Technology.

Kauffman, J. (1999). *Beginning ASP databases.* Birmingham, UK: Wrox Press.

Krumm, R. (2000). *ADO Programming for dummies.* New York: John Wiley & Sons.

MacDonald, R. (2000). *Serious ADO : Universal data access with Visual Basic.* Apress.

Martiner, W. (1999). *Building distributed applications with ADO.* New York: John Wiley & Sons.

Microsoft Universal Data Access. Retrieved February 5, 2003, from http://www.microsoft.com/data/

Morneau, K., & Batistick, J. (2001). *Active server pages.* Boston: Course Technology.

Papa, J. (2000). *Professional ADO 2.5 RDS programming with ASP 3.0.* Birmingham, UK: Wrox Press.

Roff, J. T. (2001). *ADO: ActiveX data objects.* Sebastopol, CA: O'Reilly & Associates.

Sussman, D. (2000). *Professional ADO 2.5 programming (Wrox professional guide).* Birmingham, UK: Wrox Press.

Sussman, D., & Sussman, D. (2000). *ADO 2.6, programmer's reference.* Birmingham, UK: Wrox Press.

Vaughn, W. R. (2000). *ADO Examples and best practices.* Apress.

Application Service Providers (ASPs)

Hans-Arno Jacobsen, *University of Toronto, Canada*

INTRODUCTION

The increase in network bandwidth, the growth of computing server performance, and the growing acceptance of the Internet as communication medium has given rise to a new software distribution model: application outsourcing and software leasing. Application outsourcing refers to the emerging trend of deploying applications over the Internet, rather than installing them in the local environment. Application outsourcing shifts the burden of installing, maintaining, and upgrading an application from the application user to the remote computing center, henceforth referred to as application service provider or ASP. Software leasing refers to the emerging trend of offering applications on a subscription basis, rather than through one of the traditional software licensing models. In the ASP model, system administration and application management is performed entirely by the provider. It thus becomes possible to charge a user on a pay-per-use basis, differentiable on a very fine-granular basis. This fine-grained differentiation can go as far as taking the specific functionality required by individual customers into account and metering, for computation consumed, the resources for exact billing. Thus, rather than selling a software license—giving a user “all-or-nothing” of a product—the software may be leased to the user, offering a “pay-by-need” and “pay-on-demand” model. The customer only pays for the actual functionality used and resources consumed.

A large spectrum of ASPs has become popular. Early models include hosting of database-backed Web space that offer customers solutions for hosting corporate or individual Web sites, including the access to database management systems for managing dynamic content and input. Other ASP models include the leasing of machines from computational server farms that are securely managed in reinforced buildings with high-capacity network links and power generators to guarantee uptime despite power failures. Either a customer deploys and manages its own set of machines or is assigned a dedicated set of machines on which its applications are run. Further prominent examples include online (financial) computing services, remote e-mail and document management, online accounting and billing, and Web information systems of all sorts.

The ASP-model also includes more complex application scenarios, however, such as the online access to enterprise resource planning systems (ERP), business administration applications, human resource management systems, customer relationship management systems, health care and insurance management systems, and system security management.

All these application scenarios offered as ASP solution are attractive for enterprises and individual customers alike who do not want to afford, cannot afford, or do not have the capacity to operate full-fledged stand-alone information technology (IT) systems of the described nature.

In this chapter, I provide a definition of the ASP model, describe its characteristics and facets, and discuss its implications. I begin by reviewing early developments and research trends; I then provide a detailed description of the application service-provisioning model and present a few detailed ASP examples, which lead to the identification of three ASP deployment models. I raise software-licensing issues and provide an analysis of existing ASP pricing models and strategies. I then discuss server-side implementation issues, involving a detailed description of privacy and security concerns. Finally, I draw conclusions and offer a view of how the ASP model is likely to evolve.

FROM EARLY DEVELOPMENTS TO RESEARCH TRENDS

The idea of interacting with a remote computer system across a network goes back to the mainframe era and the introduction of time-shared operating systems. Back then the driving force for this computing model was the investment-intensive mainframe computer systems that had to be maximally utilized to justify their large cost.

The 1960s can be considered the era of the mainframe. Combined with the concept of interactive computing, implemented through time-shared operating systems, succeeding the earlier batch-processing model, computing jobs had to be submitted and were processed by operators (with the final output delivered much later to the programmer) the idea of a remote computing utility, essentially today's ASP, was born. The mainframe became accessible through physically distributed dumb terminals connected with the computing utility over dedicated networks. The key difference between the remote computing model and the ASP model is that an ASP offers a fixed set of applications and services to its customers, whereas the remote computing model simply offers accessibility of the bare computing system to multiple users across the network.

In the 1970s, the minicomputer—a more affordable, smaller computing system—became the computer of choice for many companies and universities. Later on, in the 1980s, the personal computer became the lucrative choice, even spreading to the private sector. This turned the attention away from the initially popular remote computing model.

On one hand, more and more applications were developed for single-user, personal computers; on the other hand, the client–server computing model became popular. In the client–server model, a number of clients are served by a more powerful computing server. Clients and servers may either be computer systems (e.g., file server), but also may simply denote individual processes communicating with one another (e.g., database server, Web server, and application server).

In the late 1990s, due to the increasing spread and commercial acceptance of the Internet, advances in server technology, and steady increase in complexity of managing of (business) software systems, a model referred to as network computing model combined with a thin-client model became popular and set the stage for the then-emerging ASP model. Network computing again refers to accessing a powerful computing system across the network. A thin-client can range from a handheld device to a desktop but captures the notion of off-loading most computational and data management tasks to a remote computing system.

Many of the technical aspects of an application service provider have been thoroughly investigated in research; see, for example, Bhargava, King, and McQuay (1995); Czyzyk, Mesnier, and More (1997); Abel, Gaede, Taylor, and Zhou (1999); and Jacobsen, Guenther, and Riessen (2001). Business strategic and information economic questions have also been explored. For example, Marchand and Jacobsen (2001) analyzed, from an economic point of view, how the emerging ASP model may affect the profit opportunities of "traditional" independent software vendors. Two alternative economic scenarios can be envisioned, either competitively opposing application leasing and traditional licensing or combining both in a complementary fashion (see Marchand and Jacobsen for details).

The research projects exploring ASP models often go one step further than commercial ASPs do at present. Many research projects have explored more open and marketplace-oriented scenarios, in which a number of players interact. For example, the Middleware for Method Management project (see Jacobsen et al., 2001, for details) introduces the differentiation between the infrastructure provider, the data provider, the method provider, and the user. The infrastructure provider models the ASP. The data provider, a separate entity, publishes data sets that may serve the user community (e.g., historical stock quotes, geographic information, or consumer data.) The method provider publishes computational methods, which constitute algorithms from the target application domain of the marketplace (e.g., statistical analysis, numerical analysis, optimization schemes, or decision support algorithms.) The user, as in the commercial ASP model, executes published algorithms on published data. Method providers and data providers usually coincide with the application service provider in commercial systems. Interestingly, this infrastructure already recognized the need for letting individual users offer specific services for other market players to use. A similar vision, more targeting the corporate customer, is underlying the huge effort being put into the Web Services standard (World Wide Web Consortium, 2003).

Other research systems include the DecisionNet project (Bhargava et al., 1995), the NEOS service (Czyzyk et al., 1997), the SMART project (Abel et al., 1999), and MMM (Jacobsen, 2001). DecisionNet is an organized electronic market for decision support technologies. The NEOS service provides access to optimization software for use by researchers worldwide. The SMART project serves the government by assisting in county planning tasks and simplifying related administrative tasks. MMM is a middleware platform for mathematical method management that integrates various distributed mathematical software package providers, offering the user one unique access point in using the different systems.

Application Service Providers

Application service providers are third-party entities that manage, deploy, and host software-based services and applications for their customers from server farms and datacenters across wide area networks. Customers access the hosted application remotely and pay in a subscription-based manner. In essence, ASPs constitute a way for companies to outsource some or all aspects of their information technology operations, thus dramatically reducing their spending in this area. Services and applications offered by ASPs may be broadly categorized as follows:

- *Enterprise application ASPs* deliver high-end business applications, such as enterprise resource planning solutions. Customers are corporate clients who need these solutions but want to avoid investing in proper in-house installations.
- *Locally constrained ASPs* deliver a wide variety of (mostly bundled) application services in a local area, such as a portal for all the shops in a city or tourist information services for a region, including event registration, booking, and ordering features. These serve both the individual users as well as the local entity (e.g., shop or museum).
- *Specialized ASPs* deliver highly specialized applications addressing one specific function, such as news, sports and stock tickers, weather information, site indexing, or credit card validation. Customers are usually other

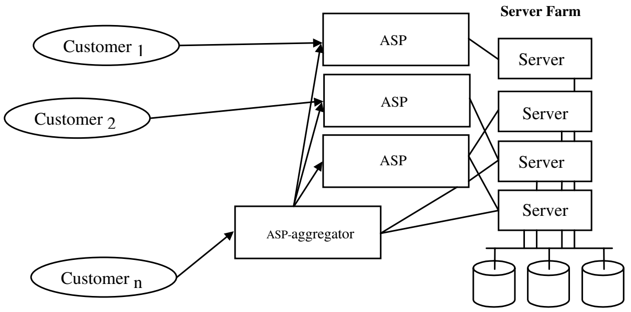


Figure 1: Application service provider architecture roles.

online service providers who bundle multiple services as one and serve one target community.

- *Vertical market ASPs* deliver applications catering to one specific industry, such as insurances, human resource management, media asset management, or health care. The customer is the corporate client.
- *Bulk-service ASPs* deliver applications for businesses in large quantities, such as e-mail, online catalog, document, or storage management.
- *ASP aggregators* combine the offerings of several ASPs and provide the user with service bundles and a single way to interact with all the aggregated ASPs.

The ASP value chain consists of many players, including the network infrastructure provider, the server farm provider, the independent software vendor, and the ASP and ASP aggregator. Figure 1 depicts a logical architecture of an ASP.

The network infrastructure provider is responsible for the network that connects customers to ASPs. The network infrastructure can be further broken down into the physical network provider, such as broadband access, phone lines, and communication infrastructure provider, and the Internet service provider, which offers services to get customers on the network. The black arrows in Figure 1 designate communication links managed by network infrastructure providers.

The server farm provider hosts the outsourced applications. The server farm consists of hundreds or more computing servers that are collectively housed. The business models for running such server farms vary. Under certain models customers bring in their own computing servers and are responsible for administering them. Other models rent a number of servers to each customer and operate them for the customer. In this case, the customer refers to the ASP operating an application. The server farm is often also referred to as a data center, because a fair amount of data management and storage is involved in most applications. In Figure 1, the server farm provider is not explicitly shown. It is responsible for the components designated as servers and as data center in the figure. The view of the server farm provider as a data center refers to ASP models in which the server farm provider takes over data backup or manages high volumes of data for the customer or for the ASP.

The independent software vendor (ISV) is responsible for the application software that is offered as outsourced solution through an ASP. Some ASPs decide to build their own software, thus avoiding the payment of license fees to the software vendor. Because the role of this player is more in the background, it is not explicitly shown in Figure 1.

The application service provider is the entity that offers the outsourced application to the customer over the network. However, the ASP must neither own the software, which it may license from an independent software vendor, nor must it own or operate the hardware, which it may lease from a server farm provider. In a further breakdown, an ASP aggregator bundles several ASPs together and offers the user one unique interface. This may be as simple as offering an ASP directory, a common log-in, and authentication, or more complex in that data can be seamlessly exchanged between the different ASPs. ASP aggregators strongly depend on open standards for accessing disparate ASPs through software integration. The emerging Web Services standards may constitute a viable solution (World Wide Web Consortium, 2003). The overall architecture of a model ASP is depicted in Figure 1, which shows the interaction of the different elements of the ASP value chain.

In the figure, an ASP is shown as a logical entity. It is associated with the software it operates for its customers and the necessary access, billing, and accounting software to run its operation. Figure 1 abstracts these functions into one component and maps them to one or more servers on the server farm. In reality, all these

Is the outsourced information highly confidential?
Does the difference of 95% versus 99% uptime make a big difference?
Is cost a major consideration and is the IT budget limited?
Is reduction of capital expenses a goal?
Dose the enterprise have remote locations that must interact with the IT system? Are there many branch offices that interact with the IT system?
Are there many (mobile) users of the IT system with a variety of client devices?
Is the lack of experienced IT staff a severe problem?
Is rapid deployment a goal?
Is the outsourced application to grow rapidly?
Are peak usage patterns expected?
Is a primary objective to focus on strategic projects rather than routine IT maintenance?

Figure 2: Factors in deciding to outsource some or all applications.

functions may operate on physically distributed computers. This achieves redundancy and fault tolerance and thus increases availability and uptime for the customer. An ASP may be a fully virtual enterprise, without any physical presence (i.e., office space).

An ASP must strategically decide to license existing applications or to build the outsourced application itself. This decision is strongly dependent on the kind of application offered. Both models can be found in practice. Existing legacy applications were never intended to be used over wide area networks by disjointed sets of users, so solutions for Web-enabling and deploying legacy applications in ASP fashion must be implemented to be used effectively in outsourcing system. Simple services, on the other hand, can easily be developed, saving the ASP expensive license and integration costs. Most ASPs are likely to use standard components, such as database management systems, in their software architectures, even for simple services. These standard components are too costly for an ASP to redevelop, so that license and integration costs are inevitable. This strategic decision by the ASP may also have an effect on customers, who may prefer to run their outsourced applications on industry-standard software. A standard ensures the customer that a switch back to an in-house operation or another ASP offering the same package can be made at any time.

The ASP model offers a number of advantages to the customer. These include a significant reduction of the total cost of ownership, because no software must be purchased and fewer IT personnel must be on site. Moreover, all IT-related tasks, such as software installation and upgrades, application maintenance, data backup, and computer system administration, are shifted to the ASP. A customer can operate with sophisticated IT applications without the huge investments in software licenses and hardware, thus drastically increasing the return of investment.

The ASP model also presents a number of disadvantages for the customer, however. These include less control over application software and the data processed, and therefore limited customization and probably less product functionality, and external dependencies on the ASP and on access to it (i.e., the network and Internet service provider.) For data storage– and data processing–intensive and nonstandard software, high switching cost is a further disadvantage to the customer. This applies equally to an application purchased for in-house use, however. Finally, the ASP model is an as yet unproven concept with little experience on either side of the relationship, little standard support, and few widely known successful applications.

For an enterprise, outsourcing part or all of its IT operation is an important strategic decision. From the previous analysis, a number of questions to guide this decision process become evident. These questions are summarized in Figure 2.

In the late 1990s, the ASP Consortium (http://www ASPstreet.com) was chartered, an industrial organization that represents the interests of ASPs and their customers. Other online resources directly related to the emerging ASP industry are the ASP Harbor (http://www Webharbour.com), ASP Island (http://www ASPIsland.com), and ASP News (http://www.ASPnews.com). These Web sites and portals are mostly commercial Web sites that collect, distribute, and sell information about ASPs.

Wireless Application Service Provider

A wireless application service provider, also known as WASP, is essentially the same as a conventional application service provider except it focuses on mobile wireless technology for service access and as a delivery mechanism. A WASP performs similar services for mobile wireless customers as the ASP does for its customers on wired lines. The wireless application service provider offers services catering to users of cellular phones, personal digital assistants, and handheld devices, and, generally, to any mobile wireless client. The service provider is more constrained in what it can offer because of the limits of the access device and great varieties in available device technology. In the business-to-consumer market, wireless application service offerings include, for example, e-mail access, unified messaging, event registration, shopping, and online banking. In the business-to-business market, wireless application services include account management and billing, backend banking, and remote sensing; in the future, this could include user location identification, system monitoring, and wireless network diagnosing. Future extensions of this model could be wireless network access providers, WASPs that offer wireless network infrastructure to customers on a pay-by-use basis in and around coffee shops, restaurants, airports, and train stations. Often, ASPs already offer wireless interaction possibilities and thus embrace both models. Because of the similarity of the WASP and ASP models, I do not discuss the WASP model further. All concepts introduced apply equally well to this kind of application service provider.

Outsourcing as a General Business Concept

Outsourcing of many business functions, from simple tasks such as bulk mailing or more complex functions such as accounting and human resource management, have been commonplace since at least the middle of the 20th century. The focus on outsourcing core IT functions and software applications is merely a special case of this broader category that has become possible because of new technology. In this chapter, the focus is on the new application service provider model enabling application outsourcing; more traditional outsourcing concepts are not covered further.

ASP EXAMPLES

In this section, I introduce a number of ASP examples. This discussion is based on existing ASP ventures. These examples have been chosen to exhibit different characteristics of ASPs that are presented more comprehensively in the following sections. The examples make no reference to specific ASP ventures, because there are too many operating ASPs, and the current ASP landscape is changing rapidly.

Income Tax ASP

The (income) tax ASP is, in the simplest instantiation, a service accessible through the Internet to complete tax forms. The yearly forms are made available on the ASP site through a browser, either based on the common HTML (hypertext markup protocol) form or based on Java applet technology. The user interface may be enhanced with features to check the entered information automatically or to guide and make the user aware of available options, essentially offering the same functionality as tax consultants or desktop tax software. The final form can then be made available as a printable document for the user to download and forward to the government tax service with supplementary information or be directly forwarded electronically to the government. In the latter case, a direct integration of the ASP with enterprise resource planning software used by the tax service can be envisioned.

Many countries are already bringing their government services, especially the revenue service, online, because it greatly facilitates the processing and distribution of the information and thus dramatically cuts costs. Moreover, private companies offer these services, combined with value-added consulting services, and forward the processed information directly to the revenue service. This example can be regarded as a successful undertaking that is changing the way people interact with their governments.

System Management ASP

The system management ASP deploys software components in the customer's computational environment that communicate autonomously with the ASP's site. In this fashion, network traffic, available disk space, system access, and, generally, any kind of computing resource activity can be closely monitored and logged. In case of a problem, or an anticipated problem, an operator is notified to take care of the problem. System intrusion detection software can be deployed in this manner as well. The local monitoring software checks for unusual system access patterns, either by forwarding network traces to the ASP's mining software or by doing the analysis locally and forwarding alert. Allen, Gabbard, and May (2003) provided a detailed discussion that investigates the outsourcing of managed IT security.

System Backup and Testing ASP

The system backup ASP offers its clients data archival services, that is, backing up disks on a defined schedule over the network, without operator intervention. Lost, overwritten, and damaged data is thus safeguarded by the ASP and can be retrieved by the client at anytime and from anywhere, over the network.

Similarly, the system-testing ASP tests a client's information system from points across the network. In many cases, a Web portal is, in manually or semiautomated fashion, subjected to loads from an outside entity; any unforeseen features, bugs, and possible errors are logged in a database and turned over to the client. Testing may include end-to-end system load measurements, monitoring of traffic, and verification of results. Because of the increase of Web-based business, this model has become a popular venture.

SOFTWARE DEPLOYMENT MODELS

The software deployment model defines where the software resides and how it is managed. In the context of application service provisioning, three models have crystallized.

Application hosting refers to the model of remotely managing a specific software package for a customer. The hosting company manages *n* applications on *n* hosts on behalf of *n* customers. Each application is given its dedicated set of resources and is physically shielded from other applications, only sharing the networking infrastructure. At one extreme of this model, the hosting company only provides the host, the network, or building infrastructure (i.e., machine rooms and physical security). Web-space hosting is a popular example of this model. Often ISP (Internet service providers), who already own appropriate data centers and server farms to start with, grow into such hosting ventures. For the hosting company it is difficult to optimally use computing resources, because a switch from one customer's system to another involves nontrivial installation steps. For example, peak load management, which shifts resources from one application to another depending on its usage pattern, is difficult to achieve for the ASP in this model. On the other hand, different hosted ventures can significantly benefit from the closeness of other hosted services, thus increasing overall efficiency of their Web-portals and decreasing perceived latency for users. For example, a big retail store whose Web site is hosted on the same server farm as an Internet advertiser will inevitably benefit from the mutual proximity. Figure 3 depicts this model logically. It shows the one-to-one correspondence between the customer and the software that the ASP operates on behalf of this customer. Each customer has a dedicated, physically separate set of computing resources. The arrows in the figure designate the network over which the customer interacts with the ASP. The figure also indicates that a single ASP manages multiple, possibly diverse, applications for different customers.

Application outsourcing refers to the model of remotely deploying one particular application, which serves many customers at the same time. This model is commonly referred to as the ASP model. Other deployment models have also been referred to as ASPs by the press, which does not differentiate carefully the various models. In this chapter, a finer grained separation is advocated. Here, in the ASP model application is offered to *n* clients. Whether

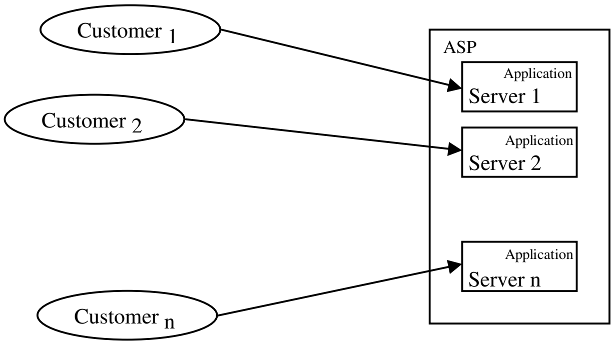


Figure 3: Application hosting deployment model.

or not this application is spread out over a number of computing nodes depends on the actual application design and implementation. The outsourced application often addresses very specific IT needs, without much customization, by many customers. Web-based e-mail services, online document management, and storage have been widely offered under such a deployment model. This approach is particularly attractive for vertical markets and applications that require little customization on a per-user basis, because in this case, the ASP has little overhead to pay. The advantage over the hosting model is that the ASP model allows computing resources to be allocated more dynamically and on demand. Server-side throughput guarantees (see Service Level Agreement later in the chapter) can thus be implemented with less investment in physical computing resources. A critical problem for this deployment model is the question of how to virtualize an application that was not intended for the use by many noninteracting customers. Standard business software has been designed for use by one customer at a time and not by *n* independent customers using it over a wide area network. Most independent software application vendors have announced their interests in this model and have started to offer their applications in such a fashion. This trend led to Web enabling of many existing applications, as well as to the redevelopment of such applications with a Web-based model in mind. Smaller and newer companies have primarily undertaken the latter.

Figure 4 depicts the application service provider deployment model. The arrows indicate network communication between customers and ASP. The difference between this model and the hosting deployment model is that here the mapping between applications managed by the ASP on behalf of its customers is not transparent. The ASP may not dedicate a physically separate server and application image for each customer; rather, the ASP may serve all customers with one application image. The customer interacts with the ASP network gateway and not the specifically dedicated resources, as in the previous model. The ASP may use less hardware to fulfill its customer needs (i.e., in Figure 4, assume *i*, the number of servers, is less than *n*, the number of servers in the hosting model).

A third model has appeared, referred to as the *application service model*. In this model, the service provider offers a service to its clients, which involves installing and maintaining software systems at the clients' site and servicing these systems and the clients' computing infrastructure across the network. Here, the service provider offers one application service to *n* clients and, additionally, manages computing systems and software at *n* client sites. Network monitoring, system administration, and system and network security constitute application services that are deployed in this manner. For the service provider this model involves great overhead, because the provider is responsible for many individually distributed, heterogeneous hardware resources. Often, a software monitor observes, at the clients' site, the state of the managed system and alerts an application service administrator preventively or in the event of problem. In a sense, this is a hybrid deployment model combining the characteristics of the previous two models. Figure 5 depicts this model logically. The components designated "ASP managed component" refer to software or hardware components that the ASP deploys on the customers site. These components monitor or control and alert the ASP in case of malfunction, emergency, or as required. These components do not exist in previous models, in which customers interact with user interface software with the remote ASP. The ASP can operate, manage, and control these components from its site over the network. These ASP-managed components are often referred to as appliances, but this term specifically refers to hardware components that are plugged into the customer's network. The mapping between customers managed and hardware required on the ASP site may correspond to any one of the previous cases.

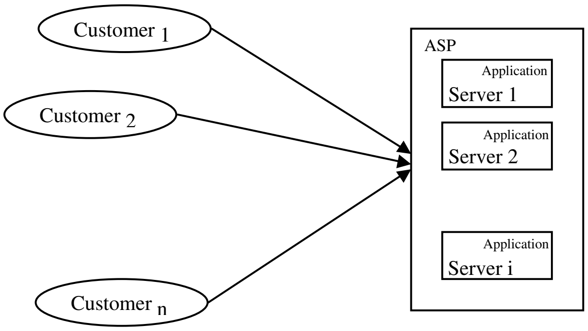


Figure 4: Application service provider deployment model.

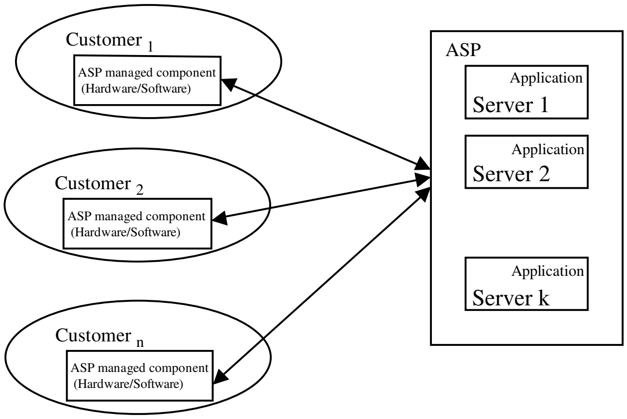


Figure 5: Application services or hybrid software deployment model.

Finally, Web services constitute services in the sense of information services, but also in the sense of human-facilitated services that are made available over the Web. In the trade press, these, too, are often referred to as application services, which is a completely different model from what has been described in this chapter thus far. The term "Web service" is in line with a set of standards (referred to as "Web Services") that is commonly used to build fully automated Web services as described here. This class of services is largely the same as the deployment of applications according to the ASP model described thus far. However, the term Web services usually refers to very specific service offers, whereas an outsourced application usually refers to a much more complex application

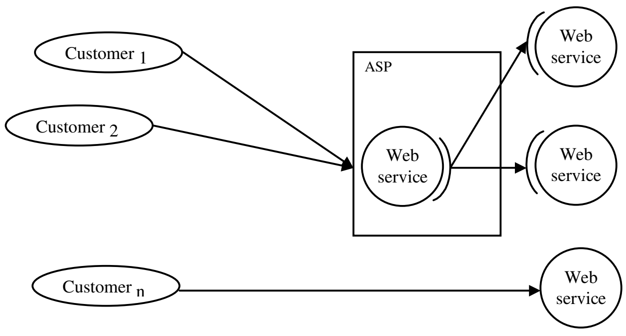


Figure 6: Web services deployment model.

offered over the network. Examples of Web services in the information technological sense are the provisioning of stock ticks, news feeds, credit card authentication, horoscopes, chat rooms, or instant messaging. Examples of Web services in the human-facilitated sense are, for example, help desks or medical and legal advice. Often a Web service is bundled, both technically and economically, into a larger Web portal offering a whole range of services to its user community. No clear-cut distinction exists between the Web service model and the application service model. Similarly, mobile services have appeared that target the mobile communication sector but essentially follow the same model.

Figure 6 shows the logical architecture of the Web service designated software deployment model. Arrows, as in the previous diagrams, designate network communication. The ASP designated component illustrates the fact that Web services are usually bundled and hardly used in isolation. The latter may also be amenable, however, for customers who directly interact with the desired Web service. The mapping between Web services and hardware to execute the service on the service deployment side depends entirely on what kind of service is offered. It is highly unlikely that a structure as found in the hosting deployment model would be propagated here. Web services may be deployed anywhere on the network and must not be physically collocated with the aggregating service. The half circles designate wrapper code that an aggregator must provide to integrate or bundle the functions of several services to offer a further new Web service. The current standardization efforts referred to as Web Services (World Wide Web Consortium, 2003) will greatly facilitate the writing of such wrapper code, because Web services conforming to this standard would expose a well-formed application programming interface.

SOFTWARE LICENSING AND DISTRIBUTION MODELS

The software licensing and distribution model defines the terms of use of the software for the customer and defines any possible obligations on the part of the software provider. Software licensing is tightly coupled with the pricing model and the service level agreement (SLA) offered by a provider. (These concepts are discussed later on in this chapter.) Four basic software licensing and distribution models broadly reflect the cases found in practice.

First, the *classical software-licensing model refers* to the case in which software is sold through a network of distribution channels to the end user, who buys and installs the software on his or her machine and uses it indefinitely. This model is not applicable in the context of application service provisioning, because the licensed application software does not reside under control by the ASP.

The *license-controlled model* refers to a refined model, whereby a customer buys and installs the software on his or her machines but is bound through a contractual agreement to pay periodic (e.g., yearly) license fees to keep using the software. In return the customer receives updates, patches, training, consulting, or new versions of the software on a regular basis. Software licenses are often designed according to the number of users working with the application (e.g., on a per-seat basis), or according to the numbers of clients interacting with an application, and on the basis of per application clients and servers deployed. Although this model already reflects a subscription-based character, it is often enforced technically by sophisticated license management software that interacts over the network with the distributor's system. This is still is not the predominant model employed by application service providers, because the customer still has most of the control over the software installed on-site.

The *leasing-based software-licensing model* refers to the model employed by application service providers. This model defines the customer's interaction with the remote application on a subscription basis and in terms of service level agreements guaranteed by the provider. The subscription mechanism defines a pricing structure based on computing resources consumed and software features used. In this model a customer interacts with an application over the network, with all management aspects of the hardware and the software being shifted to the ASP. The leased software could run entirely on the customer's machines and be managed remotely by the ASP, or part of the application could run on the customer's machine and part of it on the ASP's server farm.

Finally, a combination of the models described earlier gives rise to a further licensing and distribution model that is emerging in practice. In this model, a licensed application is offered to the client and complemented with services and extensions accessible only over the network. Examples include update distribution, library provisioning, application administration, security management, performance monitoring, and data management.

PRICING MODELS

The ASP model gives rise to the implementation of a fine-grained pricing structure that allows charging a user on a pay-per-use basis, rather than a coarse-grained structure that foresees only a limited number of prices charged for using the service. Pricing may account for the amount of system resources consumed (e.g., system interaction time, amount of data storage, CPU [computer processing unit] cycles), application functionality required, transactions executed, or simply based on a periodic or flat fee pricing model, as well as any combination of the former.

The following four pricing models can be distinguished. The flat-fee pricing model establishes a flat fee for the use of the ASP. The flat fee can be derived as a function of the size of the customer's enterprise, derived from an expected use on a per-user and per-month basis, derived from an expected transaction volume, a number of users (i.e., either named users or concurrent users), an expected usage time, or based on screen clicks. The advantage for the ASP is the simplicity to implement this model and the predictable revenue stream. The disadvantage is the invariability of the model to peak uses and customer growth (i.e., overuse). The advantage for the customer is the predictability of cost. The disadvantage for the customer is the inadaptability to underuse (i.e., paying the same price even in periods of less use).

A slightly different model, usage level-based pricing, charges a one-time setup fee and bills for certain usage levels (e.g., based on number of users using the application). Any use of the application beyond this level incurs additional charges; any use below it incurs the set charges. This model implements a predictable price structure for the customer and guarantees the ASP a usage-based remuneration, but fixed revenue if the service is underutilized by the customer. Compared with the flat-fee model, nothing changes for the customer unless the service is overused.

A usage-based pricing model bills a customer according to the resources and application features used. Implementing this kind of billing is more complicated than are the first two cases because a metering and provisioning engine must be developed for the particular hardware and software (operating system and application) used. These engines are now becoming available and are often part of new Web and application deployment infrastructures.

More difficult than metering the usage is metering of the application features used. Unless the application has been designed with this kind of deployment scenario in mind, it is difficult to offer a solution. A wrapper around the application has to be built to mediate between metering the application features and the customer interface. The disadvantage for the customer is the unpredictable price charged.

A hybrid model based on a flat fee enhanced with value-added services constitutes a further model found in practice. In addition to a flat-fee charge, the user is offered to use of value-added services that are billed according to one of the available pricing models.

Further consideration of pricing models constitutes bundling of services. This is often advocated by ASP aggregators. Several complementary services are offered as a package and are cheaper as the sum of all individual services together. A good example is the bundling of Internet access with Web hosting propagated by many ISPs today.

Furthermore, an ASP may differentiate its services by offering a variety of service-level agreements, discussed in greater detail in the next section. SLAs may account, for instance, for minimum network latency guarantees, minimum computing resource availability and throughput guarantees, and different service schedules (e.g., hotline service and data backup schedules).

Rather than selling a software license, giving a user "all or nothing" of a product, the software may be leased to the user. In contrast, under the classical software distribution model and under the software licensing model, a customer obtains the entire functionality of an application, whether or not it is actually required. Consequently, billing is much coarser grained, reflecting only the version structure of the product. A customer may, for instance, choose between a demo, a student, an advanced, and a professional version of the software, but these choices must be made up front.

Well-designed pricing models can help attract new customers and differentiate the product opening new market segments for the ASP. The right pricing model is crucial for the success of the ASP and to cover its cost. The costs for running an ASP include software license for outsourced applications, data center and network operation cost, ongoing customer support and maintenance cost, and software and infrastructure upgrade cost.

LEGAL ISSUES AND LIABILITIES

It is difficult to come up with one legal framework that determines all the responsibilities and resolves every possible conflict in application outsourcing and hosting relationships. The key problem is that the individual players may physically reside and operate in different countries bound by different legal systems but make their service available all over the Internet. Potential disputes or legal battles can severely hurt the business operation, the ASP, and the customer—as well as public opinion of the ASP model. In this section, a number of mechanisms and guidelines are discussed that can protect the customer from the ASP.

A contract with an ASP should always foresee an exit strategy that determines the conditions under which engagement with the ASP can be "legally" terminated and, in such a situation, what happens to the customer's assets (i.e., the data, business logic and process, and software that the ASP controls). This exit strategy protects the customer from a situation in which the ASP goes out of business or does not fulfill its SLA. Severe violations of SLAs may, for instance, be counteracted by a reduction in monthly payments to the ASP.

The most important rule in negotiating terms with an ASP is to retain ownership and access to all business assets outsourced to the ASP. The ASP must guarantee access to all customer data at any time and in any format requested. Copies of data should be made available regularly or, at the outer extreme, be placed in a secured location that can be physically accessed by the customer at any time. Access to data by the customer should be absolutely unconditional.

Access to critical data alone does not help to restore an IT operation once the ASP relationship has been terminated. The ASP's software is critical for processing this data. A contract with the ASP can foresee software licenses of the outsourced software for the customer and rights to operate this software in-house. This, of course, is only an option if the ASP operates standard software packages, which is not the case for all ASPs, some of which build their own solutions or integrate existing packages with value-added services. Online dispute resolution procedures have been defined by several industrial organizations in the ASP sector and for Internet-related disputes in general (WIPO Mediation and Arbitration Center, n.d.), and these may be helpful in sorting out difficulties with ASPs.

Finally, an effort toward standardization of the data formats and application programming interfaces used by ASPS may help to reduce switching cost and reduce customer lock-in, thus increasing "ASP loyalty" to a customer. This may reduce the risk of conflict in the first place.

IMPLEMENTATION ISSUES, PRIVACY, AND SECURITY CONSIDERATIONS

Service Level Agreements

An ASP customer relies on the availability of the outsourced application and the availability of network access to the application. Different customers require different classes of service, and service requirements vary among customers. For instance, customers may depend on the specific content hosted, the application or service outsourced, the time of day, or the time of year. Allocating more bandwidth and providing more server resources is a costly solution but may not resolve network congestion and server contention. On the contrary, increased resource availability attracts ever more traffic and more unrestricted use. Providing differentiated services and network access models governed by SLAs and combined with a differentiated pricing structure is the solution advocated in the service provider market today. Pricing models were discussed in the previous section. Here, I describe SLAs as related to the ASP model.

An SLA constitutes a contract between the service provider and the customer, which defines the service levels offered by the ASP to the customer. Commonly expressed in quantifiable terms, they specify levels and qualities of service that the provider guarantees and the client expects. These agreements are not unique to ASPs; many Internet service providers (ISPs) offer their customers SLAs. Also, IT departments in major enterprises have adopted the notion of defining SLAs so that services for their customers—users in other departments within the enterprise—can be quantified, justified, and ultimately compared with those provided by external ASPs. Today most service industries offer some form of SLAs.

In the ASP context, each SLA describes aspects of network and application provisioning, details the level of acceptable service, expected service, limits to the customer usage patterns, states reporting obligations, and defines conditions in case of SLA violation. An SLA contains a definition of service guarantees, which should be of an acceptable high standard and should be achievable within the pricing structure set by the service provider. Neither the customer nor the provider would not be well served by low-performance expectations, which could be achieved with ease but would not provide a sufficiently efficient and cost-effective service. Similarly, the service provider would not benefit from service levels set so high that they could not be reasonably achieved. Because network technology is rapidly changing and because the Internet's best effort service model gives rise to highly dynamic traffic patterns, SLAs cannot be defined once and for all or expressed in absolute terms. They are defined for a set period of time after which they are revisited and adapted in response to changes in technology. Because of dynamic traffic patterns, service levels are expressed as averages over time.

For ASPs service level agreements usually cover the following four areas of service levels:

1. Access to the outsourced application (i.e., network access);
2. General terms about hosting (i.e., availability of service, security, and data management);
3. Terms about the specific application or service (i.e., features supported, versions available, and upgrades administered); and
4. Customer relationship management (i.e., help desk, customization, and support).

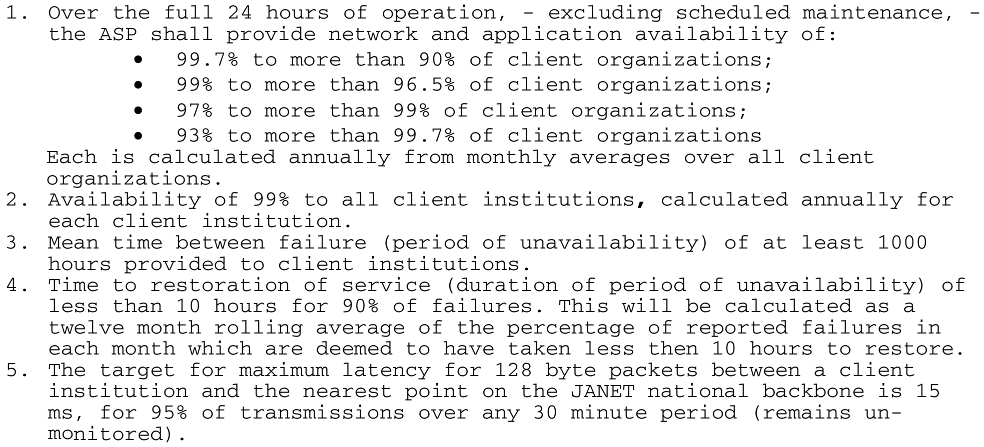
Each of these areas comprises a set of service level elements and a set of metrics to evaluate the ASP's service level guarantees and allows the customer to track its expectations and industrywide benchmarks, if available.

SLAs can be divided into technical and nontechnical service levels. A nontechnical service level guarantees a premium client a 24-hour hotline, for example, whereas a nonpremium client would only obtain a 9-to-5 weekday hotline. Various different shades of service levels are imaginable. Although nontechnical SLAs are common to all service industries, technical SLAs are more a specific characteristic of the service-provisioning model (i.e., for ISPs and ASPs alike). Technical SLAs pertain to the availability of computing and network resources that intervene in delivering the application functionality offered by the ASP to its clients. A technical SLA may guarantee the client an average throughput, an average number of transactions executed per unit of time, a bound on network bandwidth and latency. Other examples of technical SLAs are backup schedules (e.g., daily, weekly, monthly), available storage space, and security levels (e.g., encryption key strength and conformance to security standards). Example metrics that track these service levels include percentage of the time applications that are available, number of users served simultaneously, specific performance benchmarks with which actual performance is periodically compared, schedule for notification in advance of network changes that may affect users, help desk response time for various classes of problems, dial-in access availability, and usage statistics that will be provided.

Network resource guarantees, like bounds on network bandwidth and latency, are especially difficult to guarantee unless the ASP also controls the communication lines over which its service is delivered. This kind of control is unlikely in the case of Internet access. For the client, it is difficult to assess whether the promised service level has really been offered, because technical SLAs are often expressed as averages and skewed occasionally by bad user perceived latency. An extract of a real-world example of performance indicators and service levels are summarized in Figure 7.

Privacy and Security Considerations

Privacy and security is one of the primary concerns that may defeat the widespread acceptance of the ASP model. All customer application input, operational, and output



1. Over the full 24 hours of operation, - excluding scheduled maintenance, - the ASP shall provide network and application availability of:
   - 99.7% to more than 90% of client organizations;
   - 99% to more than 96.5% of client organizations;
   - 97% to more than 99% of client organizations;
   - 93% to more than 99.7% of client organizations

   Each is calculated annually from monthly averages over all client organizations.
2. Availability of 99% to all client institutions, calculated annually for each client institution.
3. Mean time between failure (period of unavailability) of at least 1000 hours provided to client institutions.
4. Time to restoration of service (duration of period of unavailability) of less than 10 hours for 90% of failures. This will be calculated as a twelve month rolling average of the percentage of reported failures in each month which are deemed to have taken less then 10 hours to restore.
5. The target for maximum latency for 128 byte packets between a client institution and the nearest point on the JANET national backbone is 15 ms, for 95% of transmissions over any 30 minute period (remains unmonitored).

Figure 7: Example of performance indicators and service levels for an Internet service provider operation.

data are available at the ASP site, potentially leaving it exposed for exploitation by the ASP, by other users, or by intruders. ASPs must implement strict security measures to

- Protect sensitive data from being stolen, corrupted, and intentionally falsified during transmission or at the remote side.
- Protect the cooperating systems from malicious use (abuse) by impersonators.
- Protect the cooperating systems from unauthorized use.
- Enforce commercial or national security concerns that may require additional steps to preserve the privacy of the data transmitted (or the encryption technology used).

To enforce these security requirements, a number of well-established techniques are available:

- Server authentication (i.e., remote site authentication ensures the client application that it is truly operating on the intended site).
- Client authentication (i.e., user authentication ensures the remote site that an authorized client is interacting).
- Integrity (i.e., noncorruption of data transferred prevents both malicious and false operation).
- Confidentiality (i.e., encrypting transferred data items prevents both malicious and false operation, as well as eavesdropping).
- Secure invocation of methods from client application to remote services, routed (i.e., delegated) through a logging facility to gather "evidence" of "who" initiated an invocation "when."
- Nonrepudiation of invoked methods to ensure liability.
- Data security to prevent sensitive or "expensive" data from being compromised at the site of computation. This may require the additional use of encryption and transformations techniques, as well as organizational means.

A detailed discussion of all mechanisms implementing these features is beyond the scope of this chapter. Except for data security, however, all these security techniques are well understood and solutions are widely deployed. The key question for an ASP user is one of trust: Why should the user entrust its sensitive and personal user or corporate data to a remote computational service (e.g., data on income, personal assets, business logic, customer information, financial data, revenue, and earnings)? Consulting groups and the trade press often note that the strongest barrier for engaging in a business relationship with an ASP is trust. Yet it is evident that for effective use of an ASP, customers must expose service input data to the ASP, where eventually it is subject to exposure, at least at the time of execution of the service's function on the data. Note that this does not refer to the risk of data being captured over the communication link. This problem is solved through cryptographic protocols that are commonplace. The data security problem is much more difficult to solve. A theoretic solution, with proven guarantees of the data remaining unknown to the ASP, is provided by Abadi and Feigenbaum (Abadi & Feigenbaum, 1990; Abadi, Feigenbaum, & Kilian, 1989) and has become know as secure circuit evaluation. Their algorithm is impractical for this scenario because it requires significant interaction between client and server to accomplish a computation. A general solution of this problem that would guarantee data security for the input, the operational, and the output data transmitted is an open research question. Approaches, such as obfuscation techniques, computing with encrypted data, computing with encrypted functions, private information retrieval, and privacy homeomorphisms are techniques that may be applied to solve this problem.

OUTLOOK

The widespread acceptance of the ASP model will depend on whether ASPs are able to ensure availability, accessibility, privacy, and security for outsourced applications and services. It will also depend on whether customers will learn to trust the model. System availability can be achieved through redundancy and replication of the service. Full accessibility of the service is strongly dependent on reliable network connections from the customer to the ASP and is less easy to guarantee by the ASP alone, unless leased communication lines are offered to access the service. Guaranteeing privacy and security of the outsourced data constitutes an open research question, with no fully satisfying solution in sight. For now, the successful ASP will offer a widely useful, highly specialized, and non-mission-critical service. Moreover, privacy guarantees and trust relationships can, for the time being, only be achieved though organizational means—for example, through the operation of ASPs by known industrial players with established brands. Because of these constraints, it is likely that the hybrid deployment model that outsources part of the software, maintains critical application components and data at the customer site, and manages these components from remote sites or, alternatively, offers additional services to complement the licensed software, will prevail.

Many enterprise resource planning (ERP) systems require significant customization to adapt the software to the business processes of the customer. Because this is a complex task to accomplish over the Web for the customer and difficult and expensive to manage on an individual basis for the ASP, it is unlikely that ASPs offering full-fledged ERP packages as leased solutions will establish themselves. If these widely accepted business software solutions establish and follow standard business process models, ASPs may be able to offer these models for a wide customer base without going through a long phase of customization.

The ASP industry is in an early stage and must still establish itself. This trend is to be expected, because the ASP model represents a paradigm shift away from the traditional application licensing and in-house management model. This emerging industry enables customers to reduce both application deployment time frames and their total cost of ownership. Although certain problems still need to be addressed, the ASP industry is successfully implementing the concept of application outsourcing and software leasing.

GLOSSARY

Appliances A prepackaged special-purpose hardware and software solution that plugs into a customer's existing IT infrastructure without much need for customization. Often, the appliance can be managed by a provider from remote.

Application hosting A computer system in a distributed environment like the Internet is often referred to as a computing host, a computing node, or simply host. Hosting refers to the provisioning of services or applications on such a computer system to make these services and applications available to authorized parties interacting in the distributed environment. Application hosting emphasizes that the hosted service is an application.

Application outsourcing The deployment of an application over a network.

Application server The system software that manages the computational transactions, business logic, and the application activation in a multitiered server configuration. The application server interacts with the database tier at one end and the client tier at the other end of the distributed application.

Application services Distinguished from applications by its smaller scope, more specific nature, and its focused functionality. An application service is seldom used in isolation; more often several application services are bundled together and offered by third-party providers that enrich their content offerings with application services.

Application service provider (ASP) The entity that manages the outsourcing of applications.

Server farm A dedicated place with many computing servers accessible through a network.

Software leasing The subscription-based offering of a software application. Often, leasing implicitly refers to a longer term relationship between the customer and the provider. Shorter term relationships are sometimes referred to as software renting, although no clear line is drawn in the context of ASPs.

Wireless application service provider (WASP) An ASP that primarily offers services to customers interacting with the ASP through wireless devices.

Web services The standard suite of protocols created by several large companies to allow applications to interoperate, discover, invoke, and integrate Web-based services across a network, primarily the Internet, is referred to as Web Services. The term Web should not be confused with the actual Web service that is created based on these standards. The Web service built in this manner is also often referred to as application service.

CROSS REFERENCES

See *Electronic Commerce and Electronic Business; Internet Literacy; Internet Navigation (Basics, Services, and Portals); Web Hosting; Web Quality of Service; Web Services.*

REFERENCES

Abadi, M., & Feigenbaum, J. (1990). Secure circuit evaluation. *Journal of Cryptology, 2,* 1–12.

Abadi, M., Feigenbaum, J., & Kilian, J. (1989). On hiding information from an oracle. *Journal of Computer and System Sciences, 39,* 21–50.

Abel, D. J., Gaede, V. J., Taylor, K. L., & Zhou, X. (1999). SMART: Towards spatial Internet marketplaces. *Geo Informatica, 3,* 141–164.

Allen, J., Gabbard, D., & May, C. Outsourcing managed security services. Technical report from the Networked Systems Survivability Program at the Software Engineering Institute. Retrieved February 2003 from http://www.cert.org/security-improvement/modules/omss/index.html

Bhargava, H. K., King, A. S., & McQuay, D. S. (1995). DecisionNet: An architecture for modeling and decision support over the World Wide Web. In T. X. Bui (Ed.), *Proceedings of the Third International Society for Decision Support Systems Conference* (Vol. II, pp. 541–550). Hong Kong: International Society for DSS.

Czyzyk, J., Mesnier, M. P., & More, J. J. (1997). *The networked-enabled optimization system (NEOS) server*

(Preprint MCS-P615—1096). Mathematics and Computer Science Division, Argonne National Laboratory.

Jacobsen, H.-A., Guenther, O., & Riessen, G. (2001). Component leasing on the World Wide Web. *NETNOMICS Journal, 2*, 191–219.

Marchand, N., & Jacobsen, H.-A. (2001). An economic model to study dependencies between independent software vendors and application service providers. *Electronic Commerce Research Journal, 1*(3), 315–334.

World Wide Web Consortium (2003). Web services specifications. Retrieved December 2002 from http://www.w3.org/2002/ws/

WIPO Mediation and Arbitration Center (n.d.). Dispute avoidance and resolution best practices for the application service provider industry. Retrieved December 2002 from http://arbiter.wipo.int/asp/report/

Authentication

Patrick McDaniel, *AT&T Labs*

AUTHENTICATION

An authentication process establishes the identity of some entity under scrutiny. For example, a traveler authenticates herself to a border guard by presenting a passport. Possession of the passport and resemblance to the attached photograph is deemed sufficient proof that the traveler is the identified person. The act of validating the passport (by checking a database of known passport serial numbers) and assessing the resemblance of the traveler is a form of authentication.

On the Internet, authentication is somewhat more complex; network entities do not typically have physical access to the parties they are authenticating. Malicious users or programs may attempt to obtain sensitive information, disrupt service, or forge data by impersonating valid entities. Distinguishing these malicious parties from valid entities is the role of authentication and is essential to network security.

Successful authentication does not imply that the authenticated entity is given access. An authorization process uses authentication, possibly with other information, to make decisions about whom to give access. For example, not all authenticated travelers will be permitted to enter the country. Other factors, such as the existence of visas, a past criminal record, and the political climate will determine which travelers are allowed to enter the country.

Although the preceding discussion focused on *entity authentication*, it is important to note that other forms of authentication exist. In particular, *message authentication* is the process by which a particular message is associated with some sending entity. This article restricts itself to entity authentication, deferring discussion of other forms to other chapters in this encyclopedia.

Meet Alice and Bob

Authentication is often illustrated through the introduction of two protagonists, Alice and Bob. In these descriptions, Alice attempts to authenticate herself to Bob. Note that Alice and Bob are often not users, but computers. For example, a computer must authenticate itself to a file server prior to being given access to its contents. Independent of whether Alice is a computer or a person, she must present evidence of her identity. Bob evaluates this evidence, commonly referred to as a credential. Alice is deemed authentic (authenticated) by Bob if the evidence is consistent with information associated with her claimed identity. The form of Alice's credential determines the strength and semantics of authentication.

The most widely used authentication credential is a password. To illustrate, UNIX passwords configured by system administrators reside in the`/etc/passwd` file. During the login process, Alice (a UNIX user) types her password into the host console. Bob (the authenticating UNIX operating system) compares the input to the known password. Bob assumes that Alice is the only entity in possession of the password. Hence, Bob deems Alice authentic because she is the only one who could present the password (credential).

Note that Bob's assertion that Alice is the only entity who could have supplied the password is not strictly accurate. Passwords are subject to *guessing* attacks. Such an attack continually retries different passwords until the authentication is successful (the correct password is guessed). Many systems combat this problem by disabling authentication (of that identity) after a threshold of failed authentication attempts. The more serious *dictionary* attack makes use of the UNIX password file itself. A salted noninvertible hash of the password is recorded in the password file. Hence, malicious parties cannot obtain the password directly from`/etc/passwd`. However, a malicious party who obtains the password file can mount a dictionary attack by comparing hashed, salted password guesses against the password file's contents. Such an attack bypasses the authentication service and, hence, is difficult to combat. Recent systems have sought to mitigate attacks on the password file by placing the password hash values in a highly restricted *shadow password* file.

Passwords are subject to more fundamental attacks. In one such attack, the adversary simply obtains the password from Alice directly. This can occur where Alice "shares" her password with others, or where she records it in some obvious place (on her PDA). Such attacks illustrate an axiom of security: A system is only as secure as the protection afforded to its secrets. In the case of authentication, failure to adequately protect credentials from misuse can result in the compromise of the system.

The definition of identity has been historically controversial. This is largely because authentication does not truly identify physical entities. It associates some secret or (presumably) unforgettable information with a virtual identity. Hence, for the purposes of authentication, any entity in possession of Alice's password is Alice. The strength of authentication is determined by the difficulty with which a malicious party can circumvent the authentication process and incorrectly assume an identity. In our above example, the strength of the authentication process is largely determined by the difficulty of guessing Alice's password. Note that other factors, such as whether Alice chooses a poor password or writes it on the front of the monitor, will determine the effectiveness of authentication. The lesson here is that authentication is as strong as the weakest link; failure to protect the password either by Alice or the host limits the effectiveness of the solution.

Authentication on the Internet is often more complex than is suggested by the previous example. Often, Alice and Bob are not physically near each other. Hence, both parties will wish to authenticate each other. In our above example, Alice will wish to ensure that she is communicating with Bob. However, no formal process is needed; because Alice is sitting at the terminal, she assumes that Bob (the host) is authentic.

On the Internet, it is not always reasonable to assume that Alice and Bob have established, or are able to establish, a relationship prior to communication. For example, consider the case where Alice is purchasing goods on the Internet. Alice goes to Bob's Web server, identifies the goods she wishes to purchase, provides her credit card information, and submits the transaction. Alice, being a cautious customer, wants to ensure that this information is only being given to Bob's Web server (i.e., authenticate the Web server). In general, however, requiring Alice to establish a direct relationship with each vendor from whom she may purchase goods is not feasible (e.g., it is not feasible to establish passwords for each Web site out of band). Enter Trent, the trusted third party. Logically, Alice appeals to Trent for authentication information relating to Bob. Trent is trusted by Alice to assert Bob's authenticity. Therefore, Bob need only establish a relationship with Trent to begin communicating with Alice. Because the number of widely used trusted third parties is small (on the order of tens), and every Web site establishes a relationship with at least one of them, Alice can authenticate virtually every vendor on the Internet.

CREDENTIALS

Authentication is performed by the evaluation of credentials supplied by the user (i.e., Alice). Such credentials can take the form of something you know (e.g., password), something you have (e.g., smartcard), or something you are (e.g., fingerprint). The credential type is specific to the authentication service and reflects some direct or indirect relationship between the user and the authentication service.

Credentials often take the form of shared secret knowledge. Users authenticate themselves by proving knowledge of the secret. In the UNIX example above, knowledge of the password is deemed sufficient evidence to prove user identity. In general, such secrets need not be statically defined passwords. For example, users in one-time password authentication systems do not present knowledge of secret text, they identify a numeric value valid only for a single authentication. Users need not present the secret directly. They need only demonstrate knowledge of it (e.g., by presenting evidence that could only be derived from it).

Secrets are often long random numbers and, thus, cannot be easily remembered by users. For example, a typical RSA private key is a 1024-digit binary number. Requiring a user to remember this number is, at the very least, unreasonable. Such information is frequently stored in a file on a user's host computer, on a PDA, or on another nonvolatile storage. The private key is used during authentication by accessing the appropriate file. However, private keys can be considered "secret knowledge" because the user presents evidence external to the authentication system (e.g., from the file system).

Credentials may also be physical objects. For example, a smartcard may be required to gain access to a host. Authenticity in these systems is inferred from possession rather than knowledge. Note that there is often a subtle difference between the knowledge- and the possession-based credentials. For example, it is often the case that a user-specific private key is stored on an authenticating smartcard. In this case, however, the user has no ability to view or modify the private key. The user can only be authenticated via the smartcard issued to the user. Hence, for the purposes of authentication, the smartcard is identity; no amount of effort can modify the identity encoded in the smartcard. Contemporary smartcards can be modified or probed. However, because such manipulation often takes considerable effort and sophistication (e.g., use of an electron microscope), such attacks are beyond the vast majority of attackers.

Biometric devices measure physical characteristics of the human body. An individual is deemed authentic if the measured aspect matches previously recorded data. The accuracy of matching determines the quality of authentication. Contemporary biometric devices include fingerprint, retina, or iris scanners and face recognition software. However, biometric devices are primarily useful only where the scanning device is trusted (i.e., under control of the authentication service). Although biometric authentication has seen limited use in the Internet, it is increasingly used to support authentication associated with physical security (i.e., governing clean-room access).

WEB AUTHENTICATION

One of the most prevalent uses of the Internet is Web browsing. Users access the Web via specialized protocols that communicate HTML and XML requests and content. The requesting user's Web browser renders received content. However, it is often necessary to restrict access to

Web content. Moreover, the interactions between the user and a Web server are often required to be private. One aspect of securing content is the use of authentication to establish the true or virtual identity of clients and Web servers.

Password-Based Web Access

Web servers initially adopted well-known technologies for user authentication. Foremost among these was the use of passwords. To illustrate the use of passwords on the Web, the following describes the configuration and use of *basic authentication* in the Apache Web server (Apache, 2002). Note that the use of basic authentication in other Web servers is largely similar.

Access to content protected by basic authentication in the Apache Web server is indirectly governed by the password file. Web-site administrators create the password file (whose location is defined by the web-site administrator) by entering user and password information using the `htpasswd` utility. It is assumed that the passwords are given to the users using an *out-of-band channel* (e.g., via e-mail, phone).

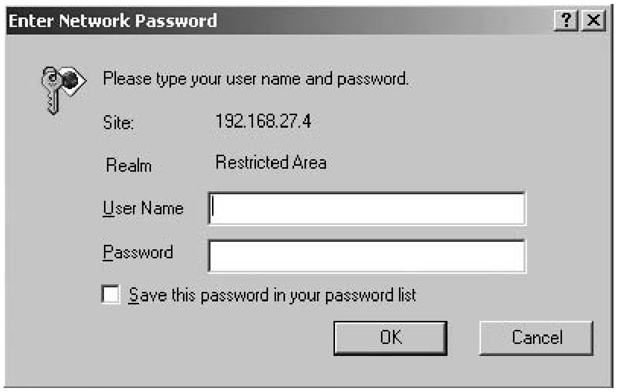
In addition to specifying passwords, the Web server must identify the subset of Web content to be password protected (e.g., a set of protected URLs). This is commonly performed by creating a `.htaccess` file in the directory to be protected. The `.htaccess` file defines the authentication type and specifies the location of the relevant password file. For example, located in the content root directory, the following `.htaccess` file restricts access to those users who are authenticated via password.

```
AuthName "Restricted Area"
AuthType Basic
AuthUserFile/var/www/webaccess
require valid-user
```

Users accessing protected content (via a browser) are presented with a password dialog (e.g., similar to the dialog depicted in Figure 1).

The user enters the appropriate username and password and, if correct, is given access to the Web content.

Because basic authentication sends passwords over the Internet in clear text, it is relatively simple to recover them by eavesdropping on the HTTP communication. Hence, basic authentication is sufficient to protect content from casual misuse but should not be used to protect valuable or sensitive data. However, as is commonly found on commercial Web sites, performing basic authentication over more secure protocols (e.g., SSL; see below) can mitigate or eliminate many of the negative properties of basic authentication.



Please type your user name and password.

Site: 192.168.27.4

Realm Restricted Area

Figure 1: Password authentication on the Web.

Many password-protected Web sites store user passwords (in encrypted form) in *cookies* the first time a user is authenticated. In these cases, the browser automatically submits the cookie to the Web site with each request. This approached eliminates the need for the user to be authenticated every time she visits the Web site. However, this convenience has a price. In most single-user operating systems, any entity using the same host will be logged in as the user. Moreover, the cookies can be easily captured and replayed back to the Web site (Fu, Sit, Smith, & Feamster, 2001).

Digest authentication uses *challenges* to mitigate the limitations of password-based authentication (Franks et al., 1999). Challenges allow authenticating parties to prove knowledge of secrets without exposing (transmitting) them. In digest authentication, Bob sends a random number (*nonce*) to Alice. Alice responds with a hash of the random number and her password. Bob uses Alice password (which only he and Alice know) to compute the correct response and compares it to the one received from Alice. Alice is deemed authentic if the computed and received responses match (because only Alice could have generated the response). Because the hash, rather than the secret, is sent, no adversary can obtain Alice's password from the response.

A number of other general-purpose services have been developed to support password maintenance. For example, RADIUS, DIAMETER, and LDAP password services have been widely deployed on the Internet. Web servers or hosts subscribing to these services defer all password maintenance and validation to a centralized service. Although each system may use different services and protocols, users see interfaces similar to those presented by basic authentication (e.g., user login above). However, passwords are maintained and validated by a centralized service, rather than by the Web server.

Single Sign-On

Basic authentication has become the predominant method of performing authentication on the web. Users often register a username and password with each retailer or service provider with which they do business. Hence, users are often faced with the difficult and error prone task of maintaining a long list of usernames and passwords. In practice, users avoid this maintenance headache by using the same passwords on all Web sites. However, this allows adversaries who gain access to the user information on one site to impersonate the user on many others.

A *single sign-on* system (SSO) defers user authentication to a single, universal authentication service. Users authenticate themselves to the SSO once per session. Subsequently, each service requiring user authentication is redirected to a SSO server that vouches for the user. Hence, the user is required to maintain only a single authentication credential (e.g., SSO password). Note that the services themselves do not possess user credentials (e.g., passwords). They simply trust the SSO to state which users are authentic.

Although single sign-on services have been used for many years (e.g., see *Kerberos*, below), the lack of universal adoption and cost of integration has made their use in Web applications highly undesirable. These difficulties have led to the creation of SSO services targeted specifically to authentication on the web. One of the most popular of these systems is the *Microsoft passport* service (Microsoft, 2002). Passport provides a single authentication service and repository of user information. Web sites and users initially negotiate secrets during the passport registration process (i.e., user passwords and Web site secret keys). In all cases, these secrets are known only to the passport servers and the registering entity.

Passport authentication proceeds as follows. Users requesting a protected Web page (i.e., a page that requires authentication) are redirected to a passport server. The user is authenticated via a passport-supplied login screen. If successful, the user is redirected back to the original Web site with an authentication *cookie* specific to that site. The cookie contains user information and site specific information encrypted with a secret key known only to the site and the passport server. The Web site decrypts and validates the received cookie contents. If successful, the user is deemed authentic and the session proceeds. Subsequent user authentication (with other sites) proceeds similarly, save that the login step is avoided. Successful completion of the initial login is noted in a session cookie stored at the user browser and presented to the passport server with later authentication requests.

Although SSO systems solve many of the problems of authentication on the Web, they are not a panacea. By definition, SSO systems introduce a single point of trust for all users in the system. Hence, ensuring that the SSO is not poorly implemented, poorly administered, or malicious is essential to its safe use. For example, passport has been shown to have several crucial flaws (Kormann & Rubin, 2000). Note that although existing Web-oriented SSO systems may be extended to support mutual authentication, the vast majority have yet to do so.

Certificates

Although passwords are appropriate for restricting access to Web content, they are not appropriate for more general Internet authentication needs. Consider the Web site for an online bookstore, *examplebooks.com*. Users wishing to purchase books from this site must be able to determine that the Web site is authentic. If not authenticated, a malicious party may impersonate *examplebooks.com* and fool the user into exposing his credit card information.

Note that most Web-enabled commercial transactions do not authenticate the user directly. The use of credit card information is deemed sufficient evidence of the user's identity. However, such evidence is typically evaluated through the credit card issuer service (e.g., checking that the credit card is valid and has not exceeded its spending limit) before the purchased goods are provided to the buyer.

The dominant technology used for Internet Web site authentication is public key certificates. Certificates provide a convenient and scalable mechanism for authentication in large, distributed environments (such as the Internet). Note that certificates are used to enable authentication of a vast array of other non-Web services. For example, certificates are often used to authenticate electronic mail messages (see *Pretty Good Privacy*, below).

Certificates are used to document an association between an identity and a cryptographic key. Keys in public key cryptography are generated in pairs: a public and a private key (Diffie & Hellman, 1976). As the name would suggest, the public key is distributed freely, and the private key is kept secret. To simplify, any data signed (using a digital signature algorithm) by the private key can be validated using the public key. A valid digital signature can be mapped to exactly one private key. Therefore, any valid signature can only be generated by some entity in possession of the private key.

Certificates are issued by *certification authorities* (CA). The CA issues a certificate by assigning an identity (e.g., the domain name of the Web site), validity dates, and the Web site's public key. The certificate is then freely distributed. A user validates a received certificate by checking the CA's digital signature. Note that most browsers are installed with a collection of CA certificates that are invariably trusted (i.e., they do not need to be validated). For example, many Web sites publish certificates issued by the Verisign CA (Verisign, 2002), whose certificate is installed with most browsers. In its most general form, a system used to distribute and validate certificates is called a *public key infrastructure*.

SSL

Introduced by Netscape in 1994, the SSL protocol uses certificates to authenticate Web content. In addition to authenticating users and Web sites, the SSL protocol negotiates an ephemeral secret key. This key is subsequently used to protect the integrity and confidentiality of all messages (e.g., by encrypting the messages sent between the Web server and the client). SSL continues to evolve. For example, the standardized and widely deployed TLS (Transport Layer Security) protocol is directly derived from SSL version 3.0.

The use of SSL is signaled to the browser and Web site through the https URL protocol identifier. For example, Alice enters the following URL to access a Web site of interest: https://www.example.com/.

In response to this request, Alice's browser will initiate an SSL handshake protocol. If the Web site is correctly authenticated via SSL, the browser will retrieve and render Web site content in a manner similar to HTTP. Authentication is achieved in SSL by validating statements signed by private keys associated with the authenticated party's public key certificate.

Figure 2 depicts the operation of the SSL authentication and key agreement process. The *SSL handshake protocol* authenticates one or both parties, negotiates the cipher-suite policy for subsequent communication (e.g., selecting cryptographic algorithms and parameters), and establishes a *master secret*. All messages occurring after

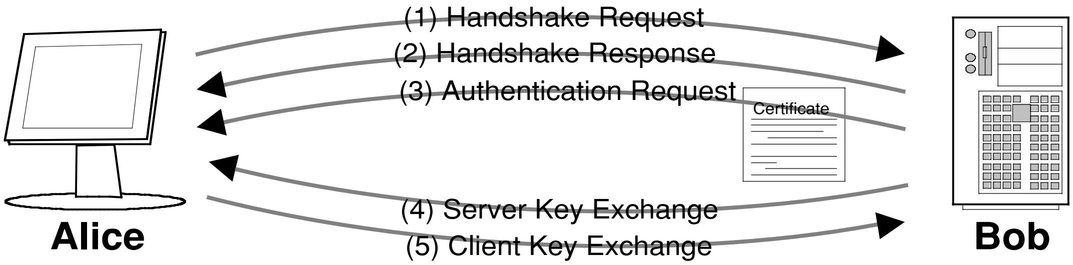


Figure 2: The SSL protocol. Alice (the client) and Bob (the server) exchange an initial handshake identifying the kind of authentication and configuration of the subsequent session security. As dictated by the authentication requirements identified in the handshake, Alice and Bob may exchange and authenticate certificates. The protocol completes by establishing a session-specific key used to secure (e.g., encrypt) later communication.

the initial handshake are protected using cryptographic keys derived from the master secret.

The handshake protocol begins with both Alice (the end-user browser) and Bob (the Web server) identifying a cipher-suite policy and session-identifying information. In the second phase, Alice and Bob exchange certificates. Note that policy will determine which entities require authentication: As dictated by policy, Alice and/or Bob will request an authenticating certificate. The certificate is validated on reception (e.g., issuance signature checked against CA's, whose certificate is installed with the browser). Note that in almost all cases, Bob will be authenticated but Alice will not. In these cases, Bob typically authenticates Alice using some external means (only when it becomes necessary). For example, online-shopping Web sites will not authenticate Alice until she expresses a desire to purchase goods, and her credit card number is used to validate her identity at the point of purchase.

Interleaved with the certificate requests and responses is the server and client key exchange. This process authenticates each side by signing information used to negotiate a session key. The signature is generated using the private key associated with the certificate of the party to be authenticated. A valid signature is deemed sufficient evidence because only an entity in possession of the private key could have generated it. Hence, signed data can be accepted as proof of authenticity. The session key is derived from the signed data, and the protocol completes with Alice and Bob sending *finished* messages.

HOST AUTHENTICATION

Most computers on the Internet provide some form of *remote access*. Remote access allows users or programs to access resources on a given computer from anywhere on the Internet. This access enables a promise of the Internet: independence from physical location. However, remote access has often been the source of many security vulnerabilities. Hence, protecting these computers from unauthorized use is essential. The means by which host authentication is performed in large part determines the degree to which an enterprise or user is protected from malicious parties lurking in the dark corners of the Internet. This section reviews the design and use of the predominant methods providing host authentication.

Remote Login

Embodying the small, isolated UNIX networks of old, *remote login* utilities allow administrators to identify the set of hosts and users who are deemed “trusted.” Trusted hosts are authenticated by source IP address, host name, and/or user name only. Hence, trusted users and hosts need not provide a user name or password.

The `rlogin` and `rsh` programs are used to access hosts. Configured by local administrators, the `/etc/hosts.equiv` file enumerates hosts/users who are trusted. Similarly, the `.rhosts` file contained in each user's home directory identifies the set of hosts trusted by an individual user. When a user connects to a remote host with a remote log-in utility, the remote log-in server (running on the accessed host) scans the `hosts.equiv` configuration file for the address and user name of the connecting host. If found, the user is deemed authentic and allowed access. If not, the `.rhosts` file of the accessing user (identified in the connection request) is scanned, and access is granted where the source address and user name is matched.

The remote access utilities do not provide strong authentication. Malicious parties may trivially forge IP addresses, DNS records, and user names (called *spoofing*). Although recent attempts have been made to address the security limitations of the IP protocol stack (e.g., IPsec, DNSsec), this information is widely accepted as untrustworthy. Remote access tools trade security for ease of access. In practice, these tools often weaken the security of network environments by providing a vulnerable authentication mechanism. Hence, the use of such tools in any environment connected to the Internet is considered extremely dangerous.

SSH

The early standards for remote access, `telnet` and `ftp`, authenticated users by UNIX password. While the means of authentication were similar to terminal log in, their use on an open network introduces new vulnerabilities. Primarily, these utilities are vulnerable to password sniffing. Such attacks passively listen in on the network for communication between the host and the remote user. Note that the physical media over which much local network communication occurs is the Ethernet. Because Ethernet is a broadcast technology, all hosts on the local network (subnet) receive every bit of transmitted data. Obviously, this approach simplifies communication eavesdropping. Although eavesdropping may be more difficult over other network media (e.g., switched networks), it is by no means impossible. Because passwords are sent in the clear (unencrypted), user-specific authentication information could be recovered. For this reason, the use of these utilities as a primary means of user access has largely been abandoned. `Ftp` is frequently used on the Web to transfer files. When used in this context, `ftp` generally operates in *anonymous* mode. `Ftp` performs no authentication in this mode, and the users are often restricted to file retrieval only.

The secure shell (SSH) (Ylonen, 1996) combats the limitations of standard tools by performing cryptographically supported host and/or user authentication. Similar to SSL, a by-product of the authentication is a cryptographic key used to obscure and protect communication between the user and remote host. SSH is not vulnerable to sniffing attacks and has been widely adopted as a replacement for the standard remote access tools.

SSH uses public key cryptography for authentication. On installation, each server host (a host allowing remote access via SSH) generates a public key pair. The public key is manually stored at each initiating host (a host from which a user will remotely connect). Note that unlike SSL, SSH uses public keys directly, rather than issued certificates. Hence, SSH authentication relies on host administrators maintaining the correct set of host keys.

SSH initiates a session in two phases. In the first phase, the server host is authenticated. The initiating host initiates the SSH session by requesting remote access. To simplify, the requesting host generates a random session key, encrypts it with a received host public key of the server, and forwards it back to the server.

The server recovers the session key using its host private key. Subsequent communication between the hosts is protected using the session key. Because only someone in possession of the host private key could have recovered the session key, the server is deemed authentic. The server transmits a short-term public key in addition to the host key. The requesting host encrypts the random value response with both keys. The use of the short-term keys prevents adversaries from recovering the content of past sessions should the host key become compromised.

The second phase of SSH session initialization authenticates the user. Dictated by the configured policy, the server will use one of the following methods to authenticate the user.

.rhosts file: As described for the *remote access* utilities above, this file simply tests whether the accessing user identifier is present in the `.rhosts` file located in the home directory of the user.

.rhosts with RSA: Similar to the above file, this requires that the accessing host be authenticated via a known and trusted RSA public key.

Password authentication: This prompts the user for a local system password. The strength of this approach is determined by the extent to which the user keeps the password private.

RSA user authentication: This works via a user-specific RSA public key. Of course, this requires that the server be configured with the public key generated for each user.

Note that it is not always feasible to obtain the public key of each host that a user will access. Host keys may change frequently (as based on an administrative policy), be compromised, or be accidentally deleted. Hence, where the remote host key is not known (and the configured policy allows it), SSH will simply transmit it during session initialization. The user is asked if the received key should be accepted. If accepted, the key is stored in the local environment and is subsequently used to authenticate the host.

Although the automated key distribution mode does provide additional protection over conventional remote access utilities (e.g., sniffing prevention), the authentication mechanism provides few guarantees. A user accepting the public key knows little about its origin (e.g., is subject to forgery, man-in-the-middle attacks, etc.). Hence, this mode may be undesirable for some environments.

One-Time Passwords

In a very different approach to combating password sniffing, the S/Key system (Haller, 1994) limits the usefulness of recovered passwords. Passwords in the S/Key system are valid only for a single authentication. Hence, a malicious party gains nothing by recovery of a previous password (e.g., via eavesdropping of a telnet log in). Although, on the surface, a one-time password approach may seem to require that the password be changed following each log in, the way in which passwords are generated alleviates the need for repeated coordination between the user and remote host.

The S/Key system establishes an ordered list of passwords. Each password is used in order and only once, then discarded. While the maintenance of the password list may seem like an unreasonable burden to place on a user, the way in which the passwords are generated makes it conceptually simple. Essentially, passwords are created such that the knowledge of a past password provides no information about future passwords. However, if one knows a secret value (called a seed value), then all passwords are easily computable. Hence, while an authentic user can supply passwords as they are needed, a malicious adversary can only supply those passwords that have been previously used (and are no longer valid).

In essence, the S/Key system allows the user to prove knowledge of the password without explicitly stating it. Over time, this relatively simple approach has been found to be extremely powerful, and it is used as the basis of many authentication services. For example, RSA's widely used SecurID combines a physical token with one-time password protocols to authenticate users (RSA, 2002).

Kerberos

The Kerberos system (Neuman & Ts'o, 1994) performs *trusted third party* authentication. In Kerberos, users, hosts, and services defer authentication to a mutually trusted key distribution center (KDC). All users implicitly trust the KDC to act in their best interest. Hence, this approach is appropriate for localized environments (e.g.,

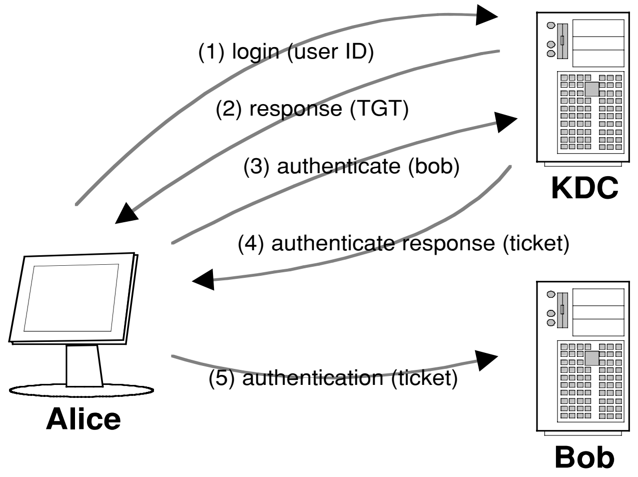


Figure 3: Kerberos authentication. Alice receives a ticket-granting ticket (TGT) after successfully logging into the Kerberos key distribution center (KDC). Alice performs mutual authentication with Bob by presenting a ticket obtained from the KDC to Bob. Note that Bob need not communicate with the KDC directly; the contents of the ticket serve as proof of Alice's identity.

campus, enterprise) but does not scale well to large, loosely coupled communities. Note that this is not an artifact of the Kerberos system but is true of any trusted third party approach; loosely coupled communities are unlikely to universally trust a single authority.

Depicted in Figure 3, the Kerberos system performs mediated authentication between Alice and Bob through a two-phase exchange with the KDC. When logging onto the system, Alice enters her user name and password.

Alice's host sends the KDC her identity. In response, the KDC sends Alice information that can only be understood by someone in possession of the password (which is encrypted with a key derived from the password). Included in this information is a ticket granting ticket (TGT) used later by Alice to initiate a session with Bob. Alice is deemed authentic because she is able to recover the TGT.

At some later point, Alice wishes to perform mutual authentication with another entity, Bob. Alice informs the KDC of this desire by identifying Bob and presenting the previously obtained TGT. Alice receives a message from the KDC containing the session key and a ticket for Bob. Encrypting the message with a key known only to Alice ensures that its contents remain confidential.

Alice then presents the ticket included in the message to Bob. Note that the ticket returned to Alice is opaque; its contents are encrypted using a key derived from Bob's password. Therefore, Bob is the only entity who can retrieve the contents of the ticket. Because later communication between Alice and Bob uses the session key (given to Alice and contained in the ticket presented to Bob), Alice is assured that Bob is authentic. Bob is assured that Alice is authentic because Bob's ticket explicitly contains Alice's identity.

One might ask why Kerberos uses a two-phase process. Over the course of a session, Alice may frequently need to authenticate a number of entities. In Kerberos, because Alice obtains a TGT at log in, later authentication can be performed automatically. Thus, the repeated authentication of users and services occurring over time does not require human intervention; Alice types in her password exactly once.

Because of its elegant design and technical maturity, the Kerberos system has been widely accepted in local environments. Historically common in UNIX environments, it has recently been introduced into other operating systems (e.g., Windows 2000, XP).

Pretty Good Privacy

As indicated by the previous discussion of trusted third parties, it is often true that two parties on the Internet will not have a direct means of performing authentication. For example, a programmer in Great Britain may not have any formal relationship with a student in California. Hence, no trusted third party exists to which both can defer authentication. A number of attempts have been made to address this problem by establishing a single public key infrastructure spanning the Internet. However, these structures require that users directly or indirectly trust CAs whose operation they know nothing about. Such assumptions are inherently dangerous and have been largely rejected by user communities.

The pretty-good-privacy (PGP) system (Zimmermann, 1994) takes advantage of informal social and organization relationships between users on the Internet. In PGP, each user creates a self-signed PGP certificate identifying a public key and identity information (e.g., e-mail address, phone number, name). Users use the key to sign the keys of those users they trust. Additionally, they obtain signatures from those users who trust them. The PGP signing process is not defined by PGP. Users commonly will exchange signatures with friends and colleagues.

The keys and signatures defined for a set of users defines a *Web of trust*. On recept of a key from a previously unknown source, an entity will make a judgment as to whether to accept the certificate based on the presence of signatures by known entities. A certificate will likely be accepted if a signature generated by a trusted party (with known and acceptable signing practices) is present. Such assessment can span multiple certificates, where signatures create trusted linkage between acceptable certificates. However, because trust is not frequently transitive, less trust is associated with long chains. PGP certificates are primarily used for electronic mail but have been extended to support a wide range of data exchange systems (e.g., Internet newsgroups).

The PGP approach and other technologies are used as the basis of S/MIME standards (Dusse, Hoffman, Ramsdell, Lundblade, & Repka, 1998). S/MIME defines protocols, data structures, and certificate management infrastructure for authentication and confidentiality of MIME (multipurpose Internet mail extensions) data. These standards are being widely adopted as a means of securing personal and enterprise e-mail.

IPsec

IPsec (Kent & Atkinson, 1998) is emerging as an important service for providing security on the Internet. IPsec is not just an authentication service but also provides a complete set of protocols and tools for securing IP-based communication. The IPsec suite of protocols provides host-to-host security within the operating system implementation of the IP protocol stack. This has the advantage of being transparent to applications running on the hosts. A disadvantage of IPsec is that it does not differentiate between users on the host. Hence, although communication passing between the hosts is secure (as determined by policy), little can be ascertained as to the true identity of users on those hosts.

The central goal of the IPsec was the construction of a general-purpose security infrastructure supporting many network environments. Hence, IPsec supports the use of an array of authentication mechanisms. IPsec authentication can be performed manually or automatically. Manually authenticated hosts share secrets distributed via administrators (i.e., configured manually at each host). Identity is inferred from knowledge of the secret. Session keys are directly or indirectly derived from the configured secret.

The Internet security association key management protocol (ISAKMP) defines an architecture for automatic authentication and key management used to support the IPsec suite of protocols. Built on ISAKMP, the Internet key exchange protocol (IKE) implements several protocols for authentication and session key negotiation. In these protocols, IKE negotiates a shared secret and policy between authenticated endpoints of an IPsec connection. The resulting IPsec Security Association (SA) records the result of IKE negotiation and is used to drive later communication between the endpoints.

The specifics of how authentication information is conveyed to a host are a matter of policy and implementation. However, in all implementations, each host must identify the keys or certificates to be used by IKE authentication. For example, Windows XP provides dialogs used to enter preshared keys. These keys are stored in the Windows registry and are later used by IKE for authentication and session key negotiation. Note that how the host stores secrets is of paramount importance. As with any security solution, users should carefully read all documentation and related security bulletins when using such interfaces.

CONCLUSION

The preceding sections described only a small fraction of a vast array of available authentication services. Given the huge number of alternatives, one might ask the question: Which one of these systems is right for my environment? The following are guidelines for the integration of an authentication service with applications and environments.

Don't try to build a custom authentication service. Designing and coding an authentication service is inherently difficult. This fact has been repeatedly demonstrated on the Internet; bugs and design flaws are occasionally found in widely deployed systems, and several custom authentication services have been broken into in a matter of hours. It is highly likely that there exists an authentication service that is appropriate for a given environment. For all of these reasons, one should use services that have been time tested.

Understand who is trusted by whom. Any authentication system should accurately reflect the trust held by all parties. For example, a system that authenticates students in a campus environment may take advantage of local authorities. In practice, such authorities are unlikely to be trusted by arbitrary endpoints in the Internet. Failure to match the trust existing in the physical world has ultimately led to the failure of many services.

Evaluate the value of the resources being protected and the strength of the surrounding security infrastructure. Authentication is only useful when used to protect access to a resource of some value. Hence, the authentication service should accurately reflect the value of the resources being protected. Moreover, the strength of the surrounding security infrastructure should be matched by the authentication service. One wants to avoid "putting a steel door in a straw house." Conversely, a weak or flawed authentication service can be used to circumvent the protection afforded by the surrounding security infrastructure.

Understand who or what is being identified. Identity can mean many things to many people. Any authentication service should model identity that is appropriate for the target domain. For many applications, it is often not necessary to map a user to a physical person or computer, but only to treat them as distinct but largely anonymous entities. Such approaches are likely to simplify authentication, and to provide opportunities for privacy protection.

Establish credentials securely. Credential establishment is often the weakest point of a security infrastructure. For example, many Web registration services establish passwords through unprotected forms (i.e., via HTTP). Malicious parties can (and do) trivially sniff such passwords and impersonate valid users. Hence, these sites are vulnerable even if every other aspect of security is correctly designed and implemented. Moreover, the limitations of many credential establishment mechanisms are often subtle. One should be careful to understand the strengths, weaknesses, and applicability of any solution to the target environment.

In the end analysis, an authentication service is one aspect of a larger framework for network security. Hence, it is a necessary to consider the many factors that contribute to the design of the security infrastructure. It is only from this larger view that the requirements, models, and design of an authentication system emerge.

GLOSSARY

Authentication The process of establishing the identity of an online entity.

Authorization The process of establishing the set of rights associated with an entity.

Certificate A digitally signed statement associating a set of attributes with a public key. Most frequently used to associate a public key with a virtual or real identity (i.e., identity certificate).

Credential Evidence used to prove identity or access rights.

Malicious party Entity on the Internet attempting to gain unauthorized access, disrupt service, or eavesdrop on sensitive communication (syn: adversary, hacker).

Secret Information only known to and accessible by a specified (and presumably small) set of entities (e.g., passwords, cryptographic keys).

Trusted third party An entity mutually trusted (typically by two end-points) to assert authenticity or authorization, or perform conflict resolution. Trusted third parties are also often used to aid in secret negotiation (e.g., cryptographic keys).

Web of trust Self-regulated certification system constructed through the creation of ad hoc relationships between members of a user community. Webs are typically defined through the exchange of user certificates and signatures within the Pretty-Good-Privacy (PGP) system.

CROSS REFERENCES

See *Biometric Authentication; Digital Identity; Digital Signatures and Electronic Signatures; Internet Security Standards; Passwords; Privacy Law; Secure Sockets Layer (SSL).*

REFERENCES

Apache (2002). Retrieved May 22, 2002, from http://httpd.apache.org/

Diffie, W., & Hellman, M. E. (1976). New directions in cryptography. *IEEE Transactions on Information Theory*, *6*, 644–654.

Dusse, S., Hoffman, P., Ramsdell, B., Lundblade, L., & Repka L. (1998). S/MIME version 2 message specification. *Internet Engineering Task Force*, RFC 2311.

Franks, J., Hallam-Baker, P., Hostetler, J., Lawrence, S., Leach, P., Luotonen, A., & Stewart, L. (1999). HTTP Authentication: Basic and digest authentication. *Internet Engineering Task Force*, RFC 2617.

Haller, N. M. (1994). The S/Key one-time password system. In *Proceedings of 1994 Internet Society Symposium on Network and Distributed System Security* (pp. 151–157). Reston, VA: Internet Society.

Kent, S., & Atkinson, R. (1998). Security architecture for the Internet protocol. *Internet Engineering Task Force*, RFC 2401.

Kevin F., Sit, E., Smith, K., & Feamster, N. (2001). Dos and don'ts of client authentication on the Web. In *10^{th} USENIX Security Symposium, 2001* (pp. 251–268). Berkeley, CA: USENIX Assocation.

Kormann, D. P., & Rubin, A. D. (2000). *Risks of the Passport Single Signon Protocol*. In *Computer Networks* (pp. 51–58). Amserdam, The Netherlands: Elsevier Science Press.

Microsoft (2002). Retrieved May 22, 2002, from http://www.passport.com/

Neuman B. C., & Ts'o T., (1994). Kerberos: An authentication service for computer networks. *IEEE Communications*, *32*(9), 33–38.

RSA (2002). Retrieved August 21, 2002, from http://www.rsasecurity.com/products/securid/

Verisign (2002). Retrieved May 22, 2002, from http://www.verisign.com/

Ylonen, T. (1996). SSH—Secure login connections over the Internet. In *Proceedings of 6th USENIX Security Symposium* (pp. 37–42). Berkeley, CA: USENIX Association.

Zimmermann, P. (1994). *PGP user's guide*. Distributed by the Massachusetts Institute of Technology, Cambridge, MA.

B

Benchmarking Internet

Vasja Vehovar, *University of Ljubljana, Slovenia*
Vesna Dolnicar, *University of Ljubljana, Slovenia*

INTRODUCTION

Benchmarking is often defined as a total quality management (TQM) tool (Gohlke, 1998). It is also one of the most recent words introduced into the lexicon of modern management (Keegan, 1998). Only since the mid-1980s have explicit benchmarking activities emerged.

With expanded benchmarking practices, a variety of professional associations have been established: the Benchmarking Exchange, Corporate Benchmarking Services, Information Systems Management Benchmarking Consortium, Telecommunications International Benchmarking Group, and International Government Benchmarking Association, to name a few. Similarly, there exist an increasing number of professional Web sites, among them the Benchmarking Exchange, Benchmarking in Europe, Public Sector Benchmarking Service, Best Practices, the Benchmarking Network, and Benchmarking. (Uniform resource locators for these organizations and Web sites are found in the Further Reading section.) A study being conducted among the members of the Benchmarking Exchange showed that the main search engine used among practitioners is Google.com, which includes almost 1 million benchmarking-related documents (Global Benchmarking Newsbrief, 2002).

The expansion also can be observed in numerous textbooks dealing either with the general notion of benchmarking or with specific benchmarking areas. Certain textbooks have already been recognized as classics (e.g., Camp, 1989). As far as periodicals are concerned, numerous professional and community newsletters arose from practical business activities, such as *eBenchmarking Newsletters* by the Benchmarking Network (n.d.), *Benchmarking News* by the European Association of Development Agencies (n.d.), and *ICOBC & Free Newsletter* by the International Council of Benchmarking Coordinators (n.d.). Benchmarking has also become a scientific issue within the field of quality management and specialized scholarly journals appeared (e.g., *Benchmarking—an International Journal and Process Management in Benchmarking*). The online academic databases searches such as *EBSCOHost—Academic Search Premier, Emerald* and *ABI/INFORM Global, Social Science Plus* show that in 2002 each of these databases already contained a minimum of 338 and a maximum of 939 papers related to benchmarking. The number of papers that relate simultaneously to benchmarking and to the Internet is considerably lower: from 3 in *Emerald* to 33 in the *EBSCOHost* database. Papers on benchmarking can be found mostly in *Internet Research, Quality Progress*, *Computerworld* and *PC Magazine*. When Longbottom (2000) reviewed approximately 500 benchmarking-related papers published between 1995 and 2000 and referenced on online academic indices (*ANBAR and Emerald*) as well as on various Internet sites, he found that the majority (80%) could be described as practical papers discussing specific aspects of benchmarking. The remaining academic papers are a mix of theory and development.

In the Web of Science, the leading science citation database, almost 2,000 benchmarking-related papers were found in 2002. Papers in this database most often refer to the issues of improving competitive advantage or to specific areas such as health care and education. Robert C. Camp, often recognized as the founder of the benchmarking concept, was the most frequently cited author in this database, with almost 700 citations. In 2002, there were almost 200 books about benchmarking available at the *Amazon.com* Web site.

Nevertheless, benchmarking predominantly relates to the business environment, although over the past few years we can observe also an increased usage in more general context. Sometimes the notion of benchmarking even appears as a synonym for any comparison based on quantitative indicators. As an example, in some studies even the simplest comparisons based on standardized statistical indicators have been labeled "benchmarking" (i.e., Courcelle & De Vil, 2001; Petrin, Sicherl, Kukar, Mesl, & Vitez, 2000). The notion of benchmarking thus has a wide range of meanings, from a specific and well-defined business practice to almost any comparison based on empirical measures. In this chapter, we use the notion of benchmarking in a broad context, although we concentrate on a specific application: the Internet. We must recognize that the Internet is a newer phenomenon than benchmarking, at least in the sense of general usage. We are thus discussing a relatively new tool (benchmarking) in a very new area (Internet). From the aspect of the scholarly investigation, the problem cannot be precisely defined. In particular, Internet-related topics can be extremely broad, as the Internet has complex consequences on a variety of subjects, from business units to specific social segments, activities, and networks, including the everyday life of the average citizen. Additional complexity arises because the Internet is automatically associated with an array of closely related issues (i.e., the new economy, new society, new business processes, new technologies) that are not clearly separated from the Internet itself. In certain areas, the Internet may have a rich and specific meaning that radically extends beyond its mere technical essence as a network of computers based on a common communication protocol.

In this chapter, we therefore understand benchmarking in its broadest sense but limit the scope of study, as much as possible, to its relation to the Internet and not to related technologies and the corresponding social and business ramifications. In particular, we limit this discussion to business entities and national comparisons.

In the following sections, we describe the notion of benchmarking in a business environment and also in a more general context, such as sectoral benchmarking and benchmarking of framework conditions. Next, we concentrate on benchmarking of Internet-related issues, particularly with regard to the performance of various countries. Key methodological problems are also discussed with specific attention to the dimension of time.

BUSINESS BENCHMARKING

A relatively sharp distinction exists between benchmarking at the company level and other types of benchmarking, which we consider extensions of the technique and discuss later in the chapter. First, we examine additional details related to the business benchmarking process.

The Concept of Benchmarking

The common denominator of various definitions of benchmarking is the concept of a "proactive, continuous process, which uses external comparisons to promote incremental improvements in products, processes, and services, which ultimately lead to competitive advantage through improved customer satisfaction, and achieving superior performance" (Camp, 1989, p. 3). The majority of authors also distinguish between *benchmarking* and *benchmarks*. The latter are measurements that gauge the performance of a function, operation, or business relative to others (i.e., Bogan & English, 1994, pp. 4–5). Similarly, Camp (1989) defined benchmark as a level of service provided, a process or a product attribute that sets the standard of excellence, which is often described as a "best-in-class" achievement. Benchmarking, in contrast to benchmarks, is the ongoing search for best practices that produce superior performance when adapted and implemented in an organization (Bogan & English, 1994). The *benchmarking process* is thus a systematic and continuous approach that involves identifying a benchmark, comparing against it, and identifying practices and procedures that will enable an organization to become the new best in class (Camp, 1989; Spendolini, 1992).

In general, two types of benchmarking definitions can be found. Some definitions are limited only to the measuring and comparing while the others focus also on implementation of change and the monitoring of results. Within this context Camp (1989, pp. 10–13) distinguished between *formal* and *working* definitions, with the latter emphasizing the decision-making component and the former relating to the measurement process alone.

As already noted, benchmarking is basically a TQM tool (Codling, 1996; Czarnecki, 1999; Gohlke, 1998). If quality management is the medicine for strengthening organizations, benchmarking is the diagnosis (Keegan, 1998, pp. 1–3). Although benchmarking readily integrates with strategic initiatives such as continuous improvement and TQM, it is also a discrete process that delivers value to the organization itself (American Productivity and Quality Center [APQC], 2002). At the extreme side, Codling (1996, pp. 24–27) did not classify benchmarking within the TQM framework at all but indicated that they are two separate processes that do not exist within a simple hierarchical relationship but are equal concepts with considerable overlap. We add that apart from TQM, benchmarking also integrates with reengineering (Bogan & English, 1994) and the Six Sigma approach (Adams Associates, 2002).

In the late 1970s, Xerox developed a well-known benchmarking project, considered a pioneer in the process (Rao et al., 1996). Xerox defined benchmarking as "a continuous process of measuring products, services, and practices against the toughest competitors or those companies recognized as industry leaders" (Camp, 1989, p. 10). Codling (1996) noted, however, that in the 1950s, well before the Xerox project, to U.K. organizations, Profit Impact of Marketing Strategy (PIMS) and the Center for Interfirm Comparison (CIFC) conducted activities that could be defined as benchmarking. PIMS and CIFC systematically gathered information on companies' performance and compared these data with those from similar businesses. Early seeds of benchmarking can be also found in the Japanese automotive industry when Toyota systematically studied U.S. manufacturing processes at General Motors, Chrysler, and Ford in the 1950s. Toyota then adopted, adapted, and improved upon their findings. All these examples confirm that companies actually used benchmarking well before 1970s, most often using the methods of site visits, reverse engineering, and competitive analysis (Rao et al., 1996).

The emphasis on formal benchmarking processes changed markedly only in 1990s, not only in the business sector, but also in regional and public sectors, particularly in Australia, New Zealand, and the United Kingdom. The initial understanding of benchmarking was rapidly extended in numerous directions. Modern benchmarking thus refers to complex procedures of evaluation, comprehension, estimation, measurement and comparison. It covers designing, processing, and interpreting of the information needed for a improved decision making. This relates not only to businesses but also to the performance of other entities, including countries. As a typical example, in the second benchmarking report comparing the performance of Belgium with other countries, Courcelle and De Vil (2001, p. 1) defined benchmarking as a continuous, systematic process for comparing performances against the best performers in the world.

Company Benchmarking

As noted earlier, benchmarking is usually a part of the quality management concept directed toward making products or services "quicker, better and cheaper" (Keegan, 1998, p. 12). The APQC (2002) suggested using benchmarking to improve profits and effectiveness, accelerate and manage change, set stretch goals, achieve breakthroughs and innovations, create a sense of urgency, overcome complacency or arrogance, see "outside the box," understand world-class performance and make better informed decisions. Within the business environment, benchmarking is most often performed in the fields of customer satisfaction, information systems, employee training, process improvement, employee recruiting, and human resources.

The literature describes many types of benchmarking processes. Camp (1995, p. 16) distinguished between four types of benchmarking: internal, competitive, functional, and generic. Similarly, Codling (1996, pp. 8–13) differentiated three types or perspectives on benchmarking: internal, external, and best practice. Bogan and English (1994, pp. 7–9) also presented three distinct types of benchmarking: process, performance, and strategic benchmarking (see also Keegan, 1998, pp. 13–16).

Benchmarking procedures are usually formalized in 4 to 12 stages (APQC, 2002; Bogan & English, 1994; Camp, 1995; Codling, 1996; Longbottom, 2000; Keegan, 1998; Spendolini, 1992). As Bogan and English (1994, p. 81) stated, the differences among benchmarking processes are often cosmetic. Most companies employ a common approach that helps them plan the project, collect and analyze data, develop insights, and implement improvement actions. Each company breaks this process into a different number of steps, however, depending on how much detail it wishes to describe at each step of the template. This does not mean that some companies exclude some steps, but in practice certain steps may naturally combine into one (Codling, 1996, p. xii). The four major stages that appear to be common to all classifications are as follows:

1. *Planning*. This step involves selection of the broad subject area to be benchmarked, defining the process, and other aspects of preparation. During the planning stage, organizations perform an internal investigation, identify potential competitors against which benchmarking may be performed, identify key performance variables, and select the most likely sources of data and the most appropriate method of data collection.
2. *Analysis*. This step involves collection of data (e.g., from public databases, professional associations, surveys and questionnaires, telephone interviews, benchmarking groups), determination of the gap between the organization's performance and that of the benchmarks, exchange of information, site visits to the benchmarked company, and observations and comparisons of process. A structured questionnaire asking for specific benchmarks, addressed to the similar or competitive business entities, is often a crucial step in collecting the data.
3. *Action*. This step involves communication throughout the organization of benchmarking results, adjustment of goals, adaptation of processes, and implementation of plans for improvement.
4. *Review*. This step involves review and repetition of the process with the goal of continuous improvement.

Another classification of the benchmarking process relates to the maturity of the company. In the early phase of the process, a company applies *diagnostic benchmarking*. The second phase is *holistic benchmarking,* in which the business as a whole is examined, identifying key areas for improvement. In the third, mature phase, the company graduates to *process benchmarking,* focusing on specific processes and chasing world-class performance (Keegan, 1998; O'Reagain & Keegan, 2000).

From these descriptions, it is clear that benchmarking activities are performed in a dialogue with competitors. As Czarnecki (1999, pp. 158, 254) pointed out, however, such a relationship does not happen overnight. Traditional barriers among competing companies must come down, and cooperation must be clearly demonstrated. Today's companies realize that to get information, they also have to give information.

Of course, for a successful implementation of change, it is important to build on the managerial foundation and culture rather than blindly adopting another organization's specific process. Edwards Deming, sometimes referred to as the father of the Japanese postwar industrial revival, illustrated this in his well-known saying that "to copy is too risky, because you don't understand why you are doing it. To *adapt* and not *adopt* is the way" (Keegan, 1998). Bogan and English (1994) pointed out that one company's effective benchmarking process design may fail at another organization with different operating concerns.

EXTENSIONS OF BENCHMARKING

Many authors (Keegan, 1998; O'Reagain & Keegan, 2000) strictly distinguish between benchmarking at the organizational (company, enterprise) level, benchmarking at the sector level, and, more generally, benchmarking of framework conditions. These extensions of benchmarking are the main focus in this section.

(Public) Sector Benchmarking

Public sector benchmarking is a natural extension of company benchmarking. Similar principles can be applied to the set of enterprises that make up an industry. Sector benchmarking thus focuses on the factors of competitiveness, which are specific to a particular industry (O'Reagain & Keegan, 2000). The usual aim here is to monitor the key factors that determine the ability of the sector to respond to continually changing international competitiveness.

During the past few decades, the notion of benchmarking extended also to a variety of nonindustrial fields, particularly to the public sector and especially to the social and welfare agencies in the health and education sector (Codling, 1996, p. 6). Of course, the goals of a public sector organization differ from those of a commercial company (O'Reagain & Keegan, 2000). For public sector organizations, benchmarking can serve as the surrogate for the competitive pressures of the market place by driving continuous improvement in value for money for taxpayers. Benchmarking can help public sector bodies to share best practices systematically with the private sector and with public bodies (e.g., government), as well as with other countries (Cabinet Office, 1999; Keegan, 1998, pp. 126–128).

A typical example of this type of benchmarking is the intra-European Union (EU) and EU—U.S. study on the performance of the national statistical offices. The comparisons of explicit benchmarks related to consideration of the time lag between data collection and the release of the economic statistics, which showed considerable lag within the EU statistical system (Statistics Sweden and Eurostat, 2001, p. 12). The study also showed, however, that in the EU international harmonization of economic statistics has been an important priority over the last decade. Further harmonization on a global level (guided by the United Nations, International Monetary Fund, and Organization for Economic Co-operation and Development [OECD]) is regarded as a much more important part of the statistical work in Europe compared with the United States, where complying with international standards has been of less importance.

Recognition that award models derived for commercial organizations can be equally applied to public sector organizations has also increased in recent years. To provide a consistent approach to assessment, some authors suggest the use of the European Foundation for Quality Management (EFQM) model for business excellence (e.g., Cabinet Office, 1999; Keegan, 1998, pp. 45–47). Keegan (1998, pp. 126–130) also mentioned "Hybrid Benchmarking," a technique that compares performance against others in both private and public sectors. Here the sources of information are similar work areas within the organization of the public sector (government departments and other public bodies) and the private sector.

Framework Conditions Benchmarking

The benchmarking method traditionally has been applied at the organizational and sector levels to evaluate the performance of the management processes, but it has been extended to the identification and the evaluation of key factors and structural conditions affecting the entire business environment. This extension is usually called the framework conditions benchmarking (Courcelle & De Vil, 2001, p. 2).

Benchmarking of framework conditions typically applies to those key elements that affect the attractiveness of a region as a place to do business. These elements can be benchmarked on a national or regional level: macroeconomic environment, taxation, labor market, education, transportation, energy, environment, research and development, foreign trade, and direct investment, as well as information and communication technology (ICT) (Courcelle & De Vil, 2001; Keegan, 1998, pp. 20–21).

Benchmarking of framework conditions therefore usually involves regions or states comparing the regulations, processes and policies that affect the business environment. Benchmarking of framework conditions usually provides an instrument for evaluating the efficiency of public policies and for identifying steps to improve them by reference to worldwide best practice (European Conference of Ministers of Transport, 2000, p. 12).

The philosophy and practice of benchmarking are roughly similar in different domains of application. However, there is an important difference in the feasibility of using results in the case of the framework conditions benchmarking, because the political power to implement changes is often lacking. Therefore one of the most important elements of the benchmarking best practice may be missing.

Benchmarking on the (Inter)national Level

In recent years, the notion of benchmarking has become extremely popular in the evaluation and comparison of countries. Theoretically, this type of benchmarking arises from benchmarking framework conditions; however, two specifics are worth noting.

Standardized comparative indicators have existed for centuries, yet the explicit label of benchmarking strongly emerged for these comparisons only with the rise of the Internet and with recent comparisons of ICT developments. Often, such notion of benchmarking for country comparisons is relatively isolated from the rich theory and practice of benchmarking. Today, we can observe national reports based on simple comparisons of indicators that are referred to as benchmarking studies; these include *Benchmarking the Framework Conditions: A Systematic Test for Belgium* (Courcelle & De Vil, 2001) and *Benchmarking Slovenia: Evaluation of Slovenia's Competitiveness, Strengths and Weaknesses* (Petrin et al., 2000). The essence of the benchmarking concept are evident in these studies because the indicators are compared with leading, comparable, or competitive countries.

Similarly, within the European Union, the notion of benchmarking has become a standard term for comparisons of the member states. Typical examples of such research are the periodic benchmark studies on the gross domestic products per capita and per employed person. In a more advanced setting, benchmarking refers to a complex process of establishing and monitoring the standardized set of indicators of the information society (e.g., Conseil de l'Union européenne, 2000).

BENCHMARKING INTERNET

Internet and Company Benchmarking

The Internet is rapidly being integrated into every facet of organizations' overall strategy and operation. Many organizations have expanded their direct-to-consumer business model, employing multiple Internet strategies to customize customer information, track customer development trends and patterns, and increase customer savings as a means to build strong relationships with their customers (Best Practices, 2000, 2001; Martin, 1999). These changes have had a major impact on the benchmarking process as well.

ICT systems are not only being benchmarked, they are also the key enablers of successful benchmarking. Current ICT systems permit users to generate, disseminate, analyze, and store vast amounts of information quickly and inexpensively. When poorly managed, however, ICT can annoy customers, slow cycle times, saddle the corporation with excessive costs, and damage productivity—all to the disadvantage of the organization (Best Practices, 2001; Bogan & English, 1994, pp. 171, 188). The benchmarks that provide the comparative insight into the role of the ICT are thus extremely important, particularly because the implementation of the ICT requires long-term strategic planning and significant investment.

The process of the Internet benchmarking on the organizational level still lacks a set of universally recognized benchmarks. Nevertheless, on the basis of several sources (e.g., Benchmark Storage Innovations, 2002; Bogan & English, 1994; Haddad, 2002; Tardugno, DiPasquale, & Matthews, 2000) we can broadly classify these benchmarks into three categories.

First, when benchmarking features and functionality, indicators usually measure the following:

- Characteristics of software and hardware (e.g., server, database, multimedia, networks, operating systems and utilities, security infrastructures, videoconference systems, corporate intranet and extranet, type of Internet connection and connection, download and upload speed)
- Purchase of new technology (e.g., share of new computers according to the number of all computers)
- Costs of technology and the organization's budget for ICT
- Software and network security administration
- Computer system performance (processing speeds, central processing unit efficiency, CD-ROM drive access speeds, performance analysis of networking and communications systems, reliability and performance modeling of software-based systems, error rates)

Second, while exploring the measures related to the use of ICT, an organization can measure the processes related to the outside environment, including its clients (customer-oriented benchmarks), or it can evaluate the use of ICTs within the organization (employee-oriented benchmarks). Customer-oriented benchmarks reflect the following:

- The extent to which ICTs have been incorporated into economic activity, such as use of the Internet in a company's transactions (i.e., electronic commerce)
- Types of e-business processes (such as Web sites with no transactions, Web-based e-commerce, electronic marketplace, etc.)
- The characteristics of strategic information technology projects (e.g., mobile or wireless commerce offerings; electronic supply chains; participation in electronic marketplaces; the organization's Web site capacity, performance, and usability; customer service and support infrastructure; creation of "localized" Web sites for customers in other countries, etc.)

Employee-oriented benchmarks typically examine the following:

- Key applications running in the organization
- Number of employees that use ICTs and the technical skills evolved
- Level of training provided for employees to use ICTs effectively
- Information system indicators reflecting organizational learning and continuous improvement
- Diffusion of telework
- New product development times
- Employee suggestion and process improvement rates
- Use of software (e.g., databases and telecommunication networks) by employees
- ICT usability

Third, when an organization explores the benefits attributed to the employment of Internet and other ICTs, it typically observes the following benchmarks:

- Increased efficiency, productivity, and performance of the organization
- Improved workstation comfort and job satisfaction
- Fewer problems in the production stage
- Broader customer base in existing and international markets
- More effective communication with customers, employees, and suppliers
- Fewer customer complaints
- Increased customer loyalty
- Better financial management
- Better integration of business processes

Measurement instruments and indicators used in the benchmarking process depend on the type of organization, its communication and business processes, the social context within which it operates, the characteristics of the employees and clients, and so on. Consequently, when evaluating features, functionality, use, and benefits of ICT, different practitioners focus on different benchmarks. In addition, most of the benchmarks described here can be measured from different perspectives; for example, a practitioner may concentrate on extent, intensity,

Table 1 Internet Benchmarks

	A[a]	B[b]	C[c]	D[d]	E[e]	F[f]	G[g]	H[h]
ICT Infrastructure								
Number of Internet hosts	x		x	x			x	x
Percentage of computers connected to the Internet			x					
Households with access to the Internet		x		x			x	
Wireless Internet access						x		
Number of Web sites								x
Price and quality of Internet connection			x				x	
Cable modem lines per 100 inhabitants				x				
Digital subscriber lines (DSL) per 100 inhabitants				x				
Network Use								
Percentage of population that (regularly) uses the Internet	x	x	x		x	x	x	x
Internet subscribers per 100 inhabitants				x			x	
Cable modem, DSL, and Internet service provider (ISP) dial-up subscribers				x				
Hours spent online per week					x			
Mobile and fixed Internet users						x		
Primary uses of the Internet					x			
Primary place of access					x			
Perception of broadband Internet access			x					
Secure Networks and Smartcards								
Number of (secure) Web servers per million inhabitants		x		x			x	x
Percentage of Internet users with security problems		x		x				
Faster Internet for Researchers and Students								
Speed of interconnections within national education networks		x						
E-commerce								
Internet access costs		x	x	x				x
Percentage of companies that buy and sell over the Internet		x		x		x		
Percentage of users ordering over the Internet				x		x		x
Business-to-consumer e-commerce transactions (% of gross domestic product)			x			x		x
Average annual e-commerce/Web spending per buyer						x		
Internet sales in the retail sector (%)				x				
Consumer Internet purchases by product				x	x	x		
Payment methods						x		
Future e-commerce plans						x		
Business intranet sophistication			x					
Online ad placement by type of site					x			
Internet advertising revenues—source comparison					x			
Domestic venture capital investment in e-commerce			x					
Competition in dot-com market			x					
Prevalence of Internet startups			x					
Use of Internet-based payment systems			x					
Sophistication of online marketing			x					
Price as barrier of e-commerce								x
Networked Learning								
Computers connected to Internet per 100 pupils		x						
Computers with high-speed connections per 100 pupils		x						
Internet access in schools		x	x					
Teachers using the Internet for noncomputing teaching		x						
Working in the Knowledge-Based Economy								
Percentage of workforce using telework		x						
Computer workers as a percentage of total employment				x				
Participation for All								
Number of public Internet points (PIAP) per 1,000 inhabitants		x						
Availability of public access to the Internet			x					

Table 1 (Continued)

	A[a]	B[b]	C[c]	D[d]	E[e]	F[f]	G[g]	H[h]
Central government Web sites that conform to the Web Accessibility Initiative (WAI)		x						
Government Online								
Percentage of basic public services available online		x	x					
Public use of government online services		x						
Percentage of online public procurement		x						
Government effectiveness in promoting the use of information and communication technology			x					
Business Internet-based interactions with government			x					
Health Online								
Percentage of health professionals with Internet access		x						
Use of different Web content by health professionals		x						

[a] European Information Technology Observatory, 10th ed. (2002).
[b] Benchmarking *e*Europe (2002).
[c] The Global IT Report (Kirkman, Cornelius, Sachs, & Schwab, 2002).
[d] Organization for Economic Co-operation and Development (2001, 2001a, 2002); Pattinson, Montagnier, & Moussiegt (2000).
[e] NUA (2002).
[f] International Data Corporation (2002c, 2002d.
[g] International Telecommunications Union (2001, 2001a).
[h] Benchmarking Belgium (Courcelle & De Vil, 2001).

quality, efficiency, mode of use, familiarity, or readiness of the certain component.

Internet and Sector Benchmarking

Sector benchmarking focuses on the factors of competitiveness, specific to a particular industry. Because of its powerful impact—which is sometimes unclear or even contradictory—it is particularly important that Internet is benchmarked within the whole sectors. The need for this practice is especially crucial in ICT-related sectors.

Typically, telecommunication companies have used benchmarking to evaluate digital versus analog technology. Benchmarks included one-time costs, maintenance costs per line, minutes of downtime per line per month, and various performance measures for processing time and failures (Bogan & English, 1994, p. 171). The Internet is being benchmarked beyond the ICT sector, however. New technologies, especially Internet-based information and service delivery, offer immense possibilities to meet a range of sector objectives. If appropriately deployed, ICT can help facilitate crucial economic and social development objectives in all sectors (World Bank Group, 2001, p. 67).

Internet and Framework Conditions Benchmarking

Framework conditions benchmarking focuses on improving the external environment in which organizations operate. One of the key elements affecting the national or regional business environment is the presence and nature of ICT. Lanvin (2002, p. xi) thus raised an important question: whether societies with different levels of development can turn the ICT revolution into an instrument that reduces the risk of marginalization and alleviates poverty. The realities in this broad and complex area require a clear assessment of how well equipped a region or country is to face the challenges of the information-driven economy (Lanvin, 2002, p. xi). So, before an action is taken, the so-called digital divide among less developed countries and the most developed countries or regions must be estimated. In other words, only when standardized indicators are available can the challenge of bridging the global digital divide be addressed.

INTERNATIONAL COMPARISONS

In this section, the key Internet indicators related to country comparisons are presented, together with their methodological specifics. The international organizations and projects that collect or present these data are briefly introduced.

Standardized Internet Benchmarks

From a technological perspective, the Internet is a global network of computers with a common communicating protocol. The corresponding social consequences of this phenomenon are extremely complex, however so we cannot avoid the benchmarks that relate not only to the Internet but also those linked to other ICT and to society. Of course, the line between the Internet and more general ICT benchmarks may be relatively vague. We limit the discussion here only to those benchmarks that are closely linked to the Internet.

In recent years, there has been a great deal of conceptual discussion about measuring Internet and the information society. The rapidly changing phenomena in this area have also challenged the process of scientific production, particularly in the social sciences, as well as the production of official statistical indicators. In last few years, however, the key Internet-related benchmarks converged to form relatively simple and commonsense standardized indicators (Table 1). This simplification corresponds to a relative loss of the enthusiasm for the so-called new economy, information society, and new business models that has recently occurred.

The quest for standardized indicators for Internet benchmarking has perhaps been strongest in the EU. In part, this is because the EU's official and ambitious goal is to surpass the United States within the next decade as the technologically most advanced society. In addition, the EU urgently needs valid comparisons among its 15 members as well as the 10 countries that will join in 2004. In addition to official EU documents regulating the standards for information society benchmarks (Benchmarking eEurope, 2002), a variety of research projects have emerged, one of the most comprehensive of which is the EU research program Statistical Indicators Benchmarking the Information Society (SIBIS, 2002). The conceptual framework for statistical measurements used by SIBIS was extensively developed for all key areas of the ICT-related phenomena—from e-security and e-commerce to e-learning, e-health, and e-government.

Currently, of course, only a small portion of the proposed indicators is being collected. Table 1 roughly summarizes only the key and most often applied benchmarks in the field of Internet-related country comparisons. In compiling the list, we sought a balance between Internet and the related ICT benchmarks and tried to avoid more general ICT indicators, such as those from the broad field of telecommunications.

The columns in Table 1 relate to the selected organizations that have published these data. Of course, the work of many other organizations was omitted because of space limitations and the scope of this chapter. Only the key international bodies and projects that systematically collect and present Internet benchmarks are listed. In addition, we also included two examples of the private companies, NUA (http://www.nua.com), which was one of the first to collect secondary data on worldwide Internet users (column E), and the International Data Corporation (IDC; http://www.idc.com), the leading global consulting agency specializing in international ICT studies (column F). In the last column (column H), the example of the benchmarks included in a typical national report (e.g., Belgium) on ICT is presented (Courcelle & De Vil, 2001). We now briefly describe the sources of the data in the Table 1.

European Information Technology Observatory (EITO)

This broad European initiative has as its objective the provision of an extensive overview of the European market for ICT within a global perspective. EITO publishes a yearbook that presents the most comprehensive and up-to-date data about the ICT market in Europe, together with the global benchmarks, particularly those related to United States and Japan (EITO, 2002). The majority of benchmarks that measure financial aspects (e.g., ICT investments) rely on data gathered by IDC. From the beginning the EITO has been strongly supported by the European Commission, Directorate General Enterprise, and Information Society, and since 1995 also by the Directorate for Science, Technology and Industry of the OECD in Paris (EITO, 2002). The annual EITO reports include the key benchmarks and also in-depth discussion of the contemporary ICT issues.

Benchmarking *e*Europe

This is the official European Union benchmarking project in the filed of ICT, begun in November 2000, when the European Council identified 23 indicators to benchmark the progress of the *eEurope Action Plan.* Indicators measure many aspects of ICTs, including e-commerce, e-government, e-security, e-education, and e-government. The facts and figures from this benchmarking program will be used to evaluate the net impact of *e*Europe and the information society, to show the current levels of activity in key areas, and to shape future policies by informing policy makers (Benchmarking *e*Europe, 2002).

The Global Information Technology Report (GITR) 2001–2002

Readiness for the Networked World is a project supported by the Information for Development Program (*info*Dev, http://www.infodev.org), a multidonor program administered by the World Bank Group (Lanvin, 2002, p. xi; World Bank Group, 2001, p. iii). At the core of the GITR is the *Networked Readiness Index,* a major comparative assessment of countries' capacity to exploit the opportunities offered by ICTs. *The Networked Readiness Index* provides a summary measure that ranks 75 countries on their relative ability to leverage their ICT networks.

Organization for Economic Co-operation and Development

The OECD groups 30 countries sharing a commitment to democratic government and the market economy. With active relationships with some 70 other countries, nongovernmental organizations, and civil societies, the OECD has a global reach. Best known for its country surveys and reviews, its work covers economic and social issues from macroeconomics to trade, education, development, and science and innovation. The OECD produces internationally agreed upon instruments to promote rules of the game in areas in which multilateral agreement is necessary for individual countries to make comparisons and progress in a global economy. Within OECD, the Statistical Analysis of Science, Technology and Industry is also conducted, together with the development of the international statistical standards for this field. Among other responsibilities, the OECD's work in this area seeks ways of examine and measure advances in science and technology and reviews recent developments in information and communication technologies (OECD, n.d.). Several internationally comparable indicators are formed within the field of the information economy, such as resources and infrastructure for the information economy, the diffusion of Internet technologies and electronic commerce, ICTs (software and hardware). The OECD also established The Committee for Information, Computer and Communications Policy (ICCP), which addresses issues arising from the digital economy, the developing global information infrastructure, and the evolution toward a global information society. In 2002, OECD published the *OECD Information Technology Outlook,* which provides a comprehensive analysis of ICTs in the economy, ICT globalization, the software sector, e-commerce, ICT skills, the digital divide, technology trends, and information technology policies.

NUA Internet Surveys

As a global resource on Internet trends, demographics, and statistics, NUA offers news and analysis updated weekly. It compiles and publishes Internet-related survey information from throughout the world. NUA is particularly known for its unique "How Many Online?" feature, which offers an estimate of the global Internet user population based on extensive examination of surveys and reports from around the world (NUA, n.d.). The value and importance of this work rapidly diminishes as more reliable and standardized indicators have begun to appear.

International Data Corporation (IDC)

IDC is a commercial company and the world's leading provider of technology intelligence, industry analysis, market data, and strategic and tactical guidance to builders, providers, and users of IT (IDC, n.d.). Thus, IDC is perhaps the most reliable global source for the number of personal computers sold in certain country or region. In addition to individual research projects and more than 300 continuous information services, IDC also provides a specific Information Society Index (ISI), which is based on four infrastructure categories: computer, information, Internet, and social infrastructures. The ISI is designed for use by governments to develop national programs that will stimulate economic and social development. It is also a tool for IT, dot-coms, and asset management and telecommunications companies with global ambitions to assess the market potential of the various regions and countries of the world (IDC, 2002a, 2002b).

International Telecommunication Union (ITU)

Headquartered in Geneva, Switzerland, ITU is an international organization within the United Nations System in which governments and the private sector coordinate global telecom networks and services. Established in 1865, ITU is the one of the world's oldest international organization. ITU's membership includes almost all countries and more than 500 private members from telecommunication, broadcasting, and IT sectors. ITU regularly publishes key telecommunication indicators, including the Internet-related benchmarks (ITU, n.d.).

Benchmarking Belgium

Benchmarking Belgium (Courcelle & De Vil, 2001) is a typical national ICT benchmarking study with the goal of comparing ICT developments in Belgium with comparable countries.

Other organizations also provide international Internet-related benchmark indicators. The indicators are usually similar to those already covered in Table 1, however. A brief listing of the most important of these follows.

The *Human Development Report*, commissioned by the United Nations Development Programme (UNDP), covers more than 100 countries annually. In 2001, the report was titled *Making New Technologies Work for Human Development*. It presents statistical cross-country comparisons that have been built up through cooperation of many organizations (e.g., several UN agencies, OECD, ITU, the World Bank). Report contains many composite indexes, such as the technology achievement index designed to capture the performance of countries in creating and diffusing technology and in building a human skills base (UNDP, 2001).

United Nations Industrial Development Organization (UNIDO) benchmarked a set of industrial performance and capability indicators and ranked 87 countries. *The Industrial Development Report 2002/2003* is intended to help policy makers, business communities, and support institutions assess and benchmark the performance of their national industries and analyze their key drivers (UNIDO, 2002).

Benchmarking is also relevant to the United Nations Educational, Scientific and Cultural Organization (UNESCO), particularly within the field of higher education. UNESCO has established Observatory of the Information Society with the objectives of raising awareness on the constant evolution of ethical, legal, and societal challenges brought about by new technologies. It aims to become a public service that provides updated information on the evolution of the information society at the national and international levels (WebWorld, 2002).

In 2001 The World Bank Group gathered data that allow comparisons for almost all existing countries available in the World Development Indicators database. Included are also indicators that measure infrastructure and access, expenditures, and business and government environment in relation to ICT.

In the United Kingdom, the Department of Trade and Industry (DTI) has sponsored research on levels of ownership, usage, and understanding of ICTs by companies of all sizes and within all sectors in benchmarked countries. The report Business in the Information Age benchmarks businesses in the United Kingdom against those in several European countries, the United States, Canada, Japan, and Australia (DTI, 2002). Also in the United Kingdom, the Office of Telecommunications (2002) issued the International Benchmarking Study of Internet Access, covering both basic dial-up access and broadband services (i.e., DSL and cable modem).

The number of institutions that publish some Internet-related measurements on international level is higher each year. It is hoped that this will also lead to accelerated establishment of standardized instruments for statistical comparisons.

Technical Measurements

The benchmarks presented in Table 1 included almost none of the performance metrics of ICT infrastructure, although they are extremely important Internet benchmarks. The technical benchmarks related to the ICT infrastructure predominantly include specific information on computers. Also relevant are the characteristics of modems and the type of Internet connection. Here, some of the most interesting benchmarks also overlap with those already outlined in the Internet and Company Benchmarking section (i.e., the type of software and hardware).

One of the central devices for the Internet technical measurement is the Internet host, where the measurements relate to corresponding speed, access, stability, and trace-route. The speed is usually expressed in the amount of information transmitted per second. Beside technical characteristics of modems and computers, the processing speed is determined by network speed between the hosts, which depends on the Internet service providers' and national communication infrastructures. In particular, the *capacity* of the total national communication links is often used as an important benchmark for the country comparisons. The access and stability are related concepts; *stability* is checked on a local level and is defined in terms of host's interruptions. The *access* stands for stability on a global level; it tells us how accessible the host is from one or more points in the Internet network. Also important are the *trace-route* reports, in which we can observe the path where data packets travel as they leave the user's computer system. More direct routes to the key international communication nodes may indicate better national infrastructure.

THE METHODOLOGICAL PROBLEMS

Of course, because of their newness, all Internet benchmarks are relatively unstable and typically face severe methodological problems. This is understandable, because these phenomena occurred relatively recently and therefore little time has been available for the discussion of methodological issues. Often, they also exhibit extremely high annual growth rates, measured in tens of percentages. In addition, the new technological improvements continuously change the nature of these phenomena and generate a permanent quest for new indicators. As a consequence, in mid-1990s this rapid development almost entirely eliminated the official statistics from this area. Instead, the private consulting agencies took the lead in ICT measurements. Thus, for example, the IDC produces many key internationally comparable data on the extent and structure of the ICT sectors.

In last few years, the efforts in official statistics and other noncommercial entities took some important steps toward compatibility. The activities within OECD particularly in Scandinavian countries, Australia, Canada, and the United States, have been particularly intensive. The United States took the lead in many respects, what was due not so much to the early Internet adoption but to early critical mass achieved in that country. In the United States, there were already millions of the Internet users by mid-1990s, a fact that many commercial organizations considered worthy of research. The U.S. government also reacted promptly, so, for example, in addition to numerous commercial measurements, official U.S. Census Bureau figures are available for business-to-customer and business-to-business sales from the end of the 1990s. The EU, in comparison, is only in the process of establishing these measurements for 2003. With respect to more sociological benchmarks, the National Telecommunications and Information Administration (2002) conducted pioneering research on the digital divide. The Pew Research Center (n.d.), a U.S. nonprofit organization, conducts important research that sets standards for sociological Internet benchmarks.

In the reminder of this section, we discuss some typical methodological problems related to the Internet benchmarks. The discussion is limited to the two most popular benchmarks in the field of Internet-related national performance: the number of Internet users and the number of Internet hosts. We believe that the methodological problems are very much typical for other indicators listed in Table 1.

Number of Internet Users

The number of Internet users heavily depends on the definition applied, an issue for which three methodological problems can be cited.

1. The Specification of Time

When defining the Internet user, usage during the last three months is often applied (NUA, 2002). Even more often, the Internet user is defined with simple self-classification, in which a question such as "Do you currently use the Internet" is asked on a survey. Experience shows that a positive answer to this question results in about 3–5% overestimation compared with questions asked among monthly users (e.g., people who claim to use the Internet on a monthly basis). Typically, usage during the last three months reveals up to a third more users compared with the category of monthly users. In the case of weekly users, which is another important benchmark, the figure shrinks to about one fifth compared with monthly Internet users. A huge variation thus exists in the number of Internet users only because of the specified frequency of usage. In addition, when asking for the Internet usage from each location separately (e.g., home, school, job), the figure increases considerably compared with asking a general question that disregards location. The timing of the survey has also a considerable impact: February figures can dramatically differ from the November figures of the same year. Unfortunately, the explicit definitions (e.g., working, timing) applied are typically not clearly stated when numbers of Internet users are published.

2. The Base and Denominator for Calculating Percentages

The number of Internet users is often observed as a share within the total population. This may be a rather unfair comparison because of populations' varying age structures and may produce artificially low figures for certain countries. Instead, often only the category 18+ is included in research, particularly in the United States. In Europe, users older than 15 years (15+) have become the standard population. The population aged 15 to 65 is also used as a basis for calculations, whereas media studies usually target the population aged 12 to 65 or 10 to 75. For a country with Internet penetration reaching about a quarter of the population, discrepancies arising from varying target populations (e.g., the basis in the denominator) vary dramatically, from the lowest Internet penetration of 20% in the population 15+ to the highest penetration of 30% in the population aged 15 to 65.

3. Internet Services Used

When asking about the Internet usage, typically only the Internet is mentioned in the survey question. Increasingly often the definition explicitly includes also the usage of the e-mail. However, here we instantly face the problem of non-Internet-based email systems. Some other definitions also include Wireless Application Protocol and other mobile Internet access methods as well as WebTV. No common international standards have been accepted. An attempt to establish such guidelines may be jeopardized with the emerging and unpredictable devices that will enable the access to the Internet. In the future, the definitions will have to become much more complex, so the potential danger for improper comparisons will also increase. The development of the standardized survey question is thus extremely important.

In addition to the these three problems, we should add that the number of Internet users is typically obtained from some representative face-to-face or telephone survey, which creates an additional and complex set of methodological problems related to the quality of survey data (sample design issues, nonresponse problems, etc.).

Another approach to estimate the number of Internet users is through models in which the number of Internet hosts and other socioeconomic parameters (i.e., educational statistics, gross domestic product) come as an input. This may be a problematic practice. A much more promising approach are so-called PC-meter measurements, in which the representative sample of Internet users is determined by installing a tracking software that records a person's Internet-related activities (e.g., Nielsen-Netratings, MediaMatrix). Despite serious methodological problems—particularly due to the non-household-PC access (i.e., business, school) and non-computer access (i.e., mobile)—this approach seems to be one of the most promising. The key advantage here is that it is not based on a survey question but on real-time observations. Another advantage is the convenience arising from the fact that the leading PC-meter companies already perform these measurements on a global level.

Number of Internet Hosts

The number of Internet hosts is perhaps the most commonly used Internet benchmarks. The reason for this is a relative easiness of its calculation and the regular frequency of these measurements. The Network Wizards (http://www.nw.com/) and Réseaux IP Européens (RIPE) (http://www.ripe.net/) are typical examples of the organizations that gather these kinds of statistics. There are severe methodological problems related to these measurements, however.

Device

The term "host" usually relates to a device that is linked to the Internet and potentially offers some content to the network. It also relates to a device with which users access the Internet. During an Internet session, each device has its Internet protocol (IP) number. The device is typically a computer; however, it can also be a modem used for a dial-up access. In the future, other devices—mobile phones, televisions, and perhaps even home appliances such as refrigerators—will also have IP numbers. National differences in the structure of those devices may post severe problems for international comparisons. Some other national specifics may also have some impact, such as a relatively large number of IP numbers partitioned on one server.

Dial-Up Modems

The most critical type of the host device is a dial-up modem, which usually serves about 100 users (e.g., households or companies) monthly. As a consequence, in each session the dial-up user connects to the Internet through a different and randomly selected modem (IP number). In countries with larger numbers of dial-up access users, the host count may underestimate the reach of the Internet.

Proxy Servers

In businesses and organizations, one computer or server may be used as the proxy host for Internet access for all computers within the local network. All the users (e.g., employees) may appear to use the same host number. Countries with a large number of such local networks may underestimate their Internet penetration.

Domain Problems

In host count statistics, all the hosts under a country's national domain are attributed to that country. The countries with restrictive domain-registration policies force their subjects to register their domains abroad, however. Consequently, a considerable number of hosts may be excluded from the national domain count. The Slovenian example is typical. Until 2003, only a company's name and trademark could receive the national domain name ".si," so up to one third of all hosts are registered under ".com," ".net," and other domains. It is true that with some additional procedures, the hosts can be reallocated to the proper country, as is typically done for the OECD. This requires additional resources, however, and is not available in the original host count data.

Technical Problems

The host count measurements are basically performed with a method "pinging" in which the computer signal is sent to a certain host number. Because of increased security protection for the local networks, the methodologies must be permanently adopted. Thus, for example, a few years ago the Network Wizards (NW) had to break the original time series of its measurements with a completely new measurement strategy. The differences between RIPE and NW are also considerable for certain countries. Local measurements can be somewhat helpful here; however, the regional or national partner may not report regularly, so a large dropout rate may result, as was often the case with RIPE data for Italy. There is also the problem of global commercial hosting, in which businesses from one country run their Web activities in the most convenient commercial space found in another country.

In the future, the host count measurement will have to upgrade measurement techniques continuously, and there will always remain certain limitations when inferring national Internet development from host count statistics.

These methodological problems related to Internet users and hosts also affect other benchmarks listed in Table 1. Thus, a general warning should be raised when using this kind of data. In particular, the methodological description must be closely observed.

Despite severe methodological problems, the national benchmarks in Table 1 offer reasonable and consistent results. Of course, with certain countries additional factors must be considered in the interpretation of data. In the future, because of the increased need for standardized, stable, and longitudinal benchmarks, we can expect that at least some of them will become standard. Another reason for this is that many phenomena have already profiled themselves and settled down in a stable and standardized form. For others, and particularly for new methods, users may have to struggle through the certain period of ambiguity during which no standardized or official indicators are available.

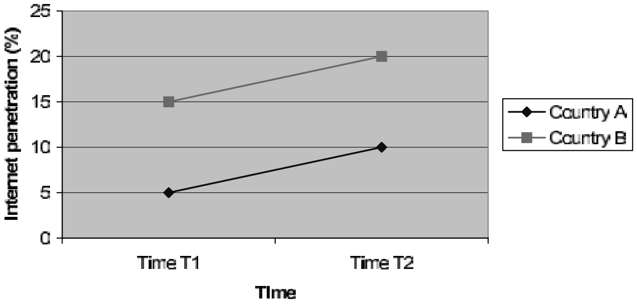


Figure 1: Internet penetration in Time 1 and in Time 2.

THE DIMENSION OF TIME

Benchmark comparisons are usually performed within time framework, so this benchmarking dimension is of great significance. Observing benchmarks through time can be extremely problematic because the straightforward comparisons of fixed benchmarks may not suffice in a rapidly changing environment.

As an example, the increase in Internet penetration from 5% in Time T1 to 10% in Time T2 for Country A demonstrates the same absolute increase in penetration as experienced by Country B with the corresponding increase from 15% (T1) to 20% (T2). In an absolute sense, one could say there had been an identical increase in Internet penetration (e.g., 5%). Similarly, the gap between the countries remains the same (e.g., 15 − 5 = 10% in time T1 and 20 − 10 = 10% in time T2).

In a relative sense, however, the increase in Country A from T1 to T2 was considerably higher:—(10 – 5)/10 = 50%, compared with (15 – 10)/15 = 33% in Country B. Similarly, the amount of the relative difference between the countries dramatically shrunk from (15 − 5)/15 = 75% at T1 compared with (20 – 10)/20 = 50% in T2. Correspondingly, at T1 Country A reached 15 – 5/15 = 33% of the Internet penetration of country B, whereas at T2 it already had reached (20 − 10)/20 = 50% of the penetration in Country B.

It is only a matter of subjective interpretation whether the differences in Internet penetration between the two countries remained the same (e.g., 5%) or decreased (e.g., Country A is reaching 50% of the penetration of Country B at T2 instead of only 33% at T1). Paradoxically, as will be shown later, the gap from T1 to T2 between these two countries most likely increased.

Of course, these differences may seem trivial because they refer to the usual statistical paradoxes, which can be dealt with a clear conceptual approach about what to benchmark together with some common sense judgment. It is much more difficult to comprehend and express the entire time dimension of the comparison in this example. The fact is that all the information regarding the time lag between the countries cannot be deduced directly from these data (Figure 1). To evaluate the entire time dimension, one would need the diffusion pattern of the Internet penetration or at least some assumptions about it. Typically, we assume that at T2 Country A will follow the pattern of Country B (Sicherl, 2001). For Figure 1 we could thus deduce, using a simple linear extrapolation, that Country A would need 2 × (T2 – T1) time units (i.e. years) to reach the penetration of the country B at T2, what is usually labeled as a time distance between the two countries.

It is also possible, however, that at T1 Country A will need, for example, 3 years to reach the penetration of Country B at T1, whereas at T2, Country A may need 5 years to reach Country B’s penetration at T2. Such an increase in lag time is expected for Internet penetration because its growth is much higher during the introductory period. Typically, much less time is needed for an increase in penetration from 5 to 10% compared with an increase from 55 to 60%. The opposite may also be true, however, as the differences in time may shrink from 3 years at T1 to 2 years at T2; it depends on the overall pattern of the Internet diffusion process.

Figure 2 demonstrates these relationships for the case of the two-dimensional presentation of the host density (the number of Internet hosts per 10,000 habitants) for Slovenia and the EU average (1995–2001). We expressed the Slovenian relative host density as the percentage of the density reached in the EU as the first dimension. The other dimension expresses the differences in terms of time distance, that is, the number of years Slovenia would need to catch up to the EU average. The method of time distance, which extrapolates the existing growth to the future, was applied here (Sicherl, 2001). In July 1995, Slovenia reached almost 40% of the EU average and in January 1997, it reached almost 90%, whereas in January 2001, it returned to 40% of the EU average. On the other hand, the corresponding time lag increased from about 1 year in 1995 to more than 3 years in 2001. The same figure for the relative benchmark (e.g., 40% in 1995 and

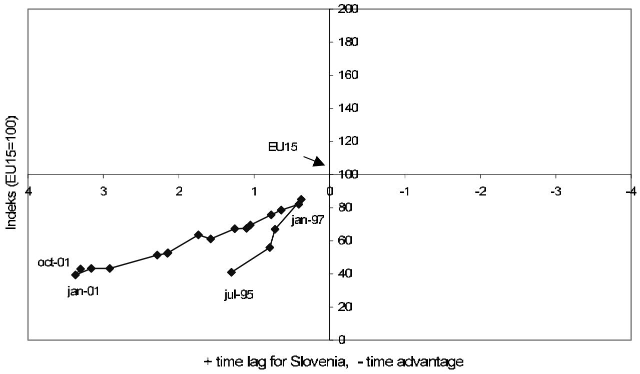


Figure 2: Host density in Slovenia and EU (1995–2002). Source: Sicherl (2001)

in 2001) has thus a dramatically different interpretation in terms of the time distance (e.g. 1 year and 3 years). The discrepancy can be explained by the fact that it was much easier to expand growth in 1995 when yearly growth rates in hosts' density were over 100% and the EU average was around 20 hosts per 10,000 habitants, compared to 2001 when the yearly growth rate were only around 10% or even stagnating, and the average of the host density was 40 hosts per 10,000 inhabitants.

Obviously, the Internet benchmarks should be observed within the framework of changing penetration patterns. Any benchmark that relies only on the comparisons of absolute or relative achievements may not be exhaustive in explaining the phenomena. It can be even directly misleading. This example illustrates that benchmark researchers must take the time dimension into careful consideration.

CONCLUSION

The basic concept of benchmarking relates to comparisons of performance indicators with a common reference point. Historically, such comparisons have been performed since the time of the ancient Egyptians. The systematic collection of the benchmarks also existed from the early days of the competitive economy, when companies compared their business practices with those of competitors. The explicit notion of benchmarking arose only in the late 1970s with the pioneering work of Xerox, however, and interest in the field exploded in the 1990s. Today benchmarking is an established discipline with professional associations, awards, codes of conduct, conferences, journals, and textbooks, and companies around the world are involved in the practice.

There are no doubts that modern benchmarking arose from a business environment where all the basic methodology and the standard procedures were developed. However, during past years the notion of benchmarking has expanded to sector benchmarking as well as to the governmental and nonprofit sector. In last few years it has also become popular for the national comparisons in the field of ICT. A number of international studies have been labeled as benchmarking, although little benchmarking theory was actually applied (Courcelle & De Vil, 2001; Petrin et al., 2000). The EU adopted benchmarking for ICT comparisons of member and candidate nations in a formal manner. In this case, statistical data are used for systematic year-by-year comparisons according to 23 Internet benchmarks.

The speed of changes in the field of ICT creates severe methodological problems for the Internet benchmarks. With the dramatic rise of the Internet in mid-90s only private companies had sufficient flexibility to provide up-to-date ICT indicators. As a consequence, even today, for the ICT international comparisons the data from private agencies are often used. In particular, this holds true for the scope and structure of the ICT spending. Only in recent years have the official statistics and other international bodies recovered from this lag and presented their own methodological outlines. Here, the work within EU and particularly within the OECD should be emphasized.

The contemporary Internet indicators used for the international comparisons of the countries' performance have stabilized only in recent years. After many theoretical discussions about the complexity of the information society, relatively simple indicators became the standards for the national ICT benchmarking. Among the key indicators in this field are the Internet penetration, the host density, and the share of Internet transactions among all commercial transactions of consumers and companies as well as within the government–citizen relations.

GLOSSARY

Benchmark A reference point, or a unit of measurement, for making comparisons. A benchmark is a criterion for success, an indicator of the extent to which an organization achieves the targets and goals defined for it.

Benchmarking A process whereby a group of organizations, usually in the same or similar domains, compare their performance on a number of indicators. The aim of the exercise is for participants to learn from each other and to identify good practice with a view toward improving performance in the long run.

CROSS REFERENCES

See *Developing Nations; Feasibility of Global E-business Projects; Global Issues; Information Quality in Internet and E-business Environments; Internet Literacy; Internet Navigation (Basics, Services, and Portals); Web Quality of Service.*

REFERENCES

Adams Associates (2002). *Six Sigma plus: Black belt training*. Retrieved June 22, 2002, from http://www.adamssixsigma.com

American Productivity and Quality Center (n.d.). Retrieved May 28, 2002, from http://www.apqc.org

Benchmark Storage Innovations (2002). *Global IT Strategies 2001*. Retrieved June 22, 2002, from http://www.informationweek.com/benchmark/globalIT.htm

Benchmarking eEurope (2002). Retrieved June 22, 2002, from http://europa.eu.int/information_society/eeurope/benchmarking/index_en.htm

Benchmarking Exchange—Benchnet (2002). *Benchmarking management report*. Retrieved June 22, 2002, from http://66.124.245.170/surveys/bmsurvey/results.cfm?CFID = 199829&CFTOKEN = 38067072

Benchmarking Network (n.d.). *eBenchmarking newsletter* (TBE newsletter archives). Retrieved June 22, 2002, from http: // 66.124.245.170 / TBE_Members2 / newsletters/index.cfm

Best Practices (2000). *BestPracticeDatabase.com: Internet & e-business*. Retrieved May 28, 2002, from http://www.bestpracticedatabase.com/subjects/internet_ebusiness.htm

Best Practices (2001). *Online report summary. Driving business through the internet: Web-based sales,* *marketing and service*. Retrieved May 28, 2002, from http://www.benchmarkingreports.com/salesandmarketing/sm140_ebusiness.asp

Bogan, C. E., & English, M. J. (1994). *Benchmarking for best practices: Winning through innovative adaptation*. New York: McGraw-Hill.

Cabinet Office (U.K.) (1999). *Public sector excellence programme*. Retrieved June 11, 2002, from http://www.cabinet-office.gov.uk/eeg/1999/benchmarking.htm

Camp, R. C. (1995). *Business process benchmarking: Finding and implementing best practices*. Milwaukee, WI: ASQ Quality Press.

Camp, R. C. (1989). *Benchmarking: The search for industry best practices that lead to superior performance*. Milwaukee, WI: ASQC Quality Press.

Codling, S. (1996). *Best practice benchmarking: An international perspective*. Houston, TX: Gulf.

Conseil de l'Union européenne (2000). *List of eEurope Benchmarking indicators*. Retrieved February 12, 2002, from http://europa.eu.int/information_society/eeurope/benchmarking/indicator_list.pdf

Courcelle, C., & De Vil, G. (2001). *Benchmarking the framework conditions: A systematic test for Belgium. Federal Planning Bureau: Economic analysis and forecasts*. Retrieved February 12, 2002, from http://www.plan.be/en/bench/index.htm

Czarnecki, M. T. (1999). *Managing by measuring: How to improve your organization's performance through effective benchmarking*. Houston, TX: Benchmarking Network.

Department of Trade and Industry (2002). *Business in the information age*. Retrieved October 22, 2002, from http://www.ukonlineforbusiness.gov.uk/main/resources/publication-htm/bench2001.htm

European Association of Development Agencies (n.d.). *Benchmarking News*. Retrieved June 22, 2002, from http://www.eurada.org/News/Benchmarking/English/ebenchtable.htm

European Conference of Ministers of Transport (2000). *Transport benchmarking: Methodologies, applications & data needs*. Paris: OECD Publications Service.

European Information Technology Observatory, 10th ed. (2002). Frankfurt am Main: Author. Retrieved June 25, 2002, from http://www.eito.com/start.html

Global Benchmarking Newsbrief (2002). *Benchmarking update*. Retrieved from June 16, 2002, from http://66.124.245.170/TBE_Members2/newsletters/bupdate 0601.cfm

Gohlke, A. (1998). *Benchmarking basics for librarians*. Retrieved June 23, 2002, from http://www.sla.org/division/dmil/mlw97/gohlke/sld001.htm

International Data Corporation (n.d.). *About IDC*. Retrieved June 20, 2002, from http://www.idc.com/en_US/st/aboutIDC.jhtml

International Data Corporation (2002a). *IDC/World Times information society index: The future of the information society*. Retrieved June 20, 2002, from http://www.idc.com/getdoc.jhtml?containerId = 24821

International Data Corporation (2002b). *Sweden remains the world's dominant information economy while the United States slips, according to the 2001 IDC/World Times Information Society Index*. Retrieved June 20, 2002, from http://www.idc.com/getdoc.jhtml?containerId = pr50236

International Council of Benchmarking Coordinators (n.d.). *ICOBC & free newsletter*. Retrieved June 22, 2002, from http://www.icobc.com

International Telecommunications Union (n.d.). *ITU overview—contents*. Retrieved June 25, 2002, from http://www.itu.int/aboutitu/overview/index.html

International Telecommunications Union (2001a). *Internet indicators*. Retrieved June 25, 2002, from http://www.itu.int/ITU-D/ict/statistics/at_glance/Internet01.pdf

International Telecommunications Union (2001b). *ITU telecommunication indicators update*. Retrieved June 25, 2002, from http://www.itu.int/ITU-D/ict/update/pdf/Update_1_01.pdf

Haddad, C. J. (2002). *Managing technological change. A strategic partnership approach*. Thousand Oaks, CA: Sage.

Keegan, R. (1998). *Benchmarking facts: A European perspective*. Dublin: European Company Benchmarking Forum.

Kirkman, G. S., Cornelius, P. K., Sachs, J. D., & Schwab, K. (Eds.). (2002). *The global information technology report 2001–2002: Readiness for the networked world*. New York: Oxford University Press.

Lanvin, B. (2002). Foreword. In G. S. Kirkman, P. K. Cornelius, J. D. Sachs, & K. Schwab (Ed.), *The global information technology report 2001–2002: Readiness for the networked world* (pp. xi–xii). New York: Oxford University Press.

Longbottom, D. (2000). Benchmarking in the UK: An empirical study of practitioners and academics. *Benchmarking: An International Journal, 7,* 98–117.

Martin, T. (1999). Extending the direct-to-consumer model through the internet. *GBC conference presentations*. Retrieved May 28, 2002, from http://www.globalbenchmarking.com/meetings/presentationdetails.asp?uniqueid = 54

National Telecommunications and Information Administration (2002). *Americans in the information age: Falling through the net*. Retrieved October 20, 2002, from http://www.ntia.doc.gov/ntiahome/digitaldivide

NUA (n.d.). *About Nua.com*. Retrieved June 20, 2002, from http://www.nua.ie/surveys/about/index.html

NUA (2002). *NUA Analysis*. Retrieved June 20, 2002, from http://www.nua.com/surveys/analysis/graphs_charts/index.html

Office of Telecommunications (2002). *International benchmarking study of Internet access*. Retrieved October 20, 2002, from http://www.oftel.gov.uk/publications/research/2002/benchint0602.pdf

O'Reagain, S., & Keegan, R. (2000). *Benchmarking explained. Benchmarking in Europe—working together to build competitiveness*. Retrieved February 12, 2002, from http://www.benchmarking-in-europe.com/library/archive_material/articles_publications/archive_psi_articles/explained.htm

Organization for Economic Co-operation and Development (n.d.). Home page. Retrieved June 20, 2002, from http://www.oecd.org

Organization for Economic Co-operation and Development (2001a). *Business to consumer electronic commerce: An update on the statistics*. Retrieved June 20, 2002, from http://www.oecd.org/pdf/M00018000/M00018264.pdf

Organization for Economic Co-operation and Development (2001b). *The latest official statistics on electronic commerce: A focus on consumers' Internet transactions*. Retrieved June 20, 2002, from http://www.oecd.org/pdf/M00027000/M00027669.pdf

Organization for Economic Co-operation and Development (2002). *OECD information technology outlook*. Retrieved June 20, 2002, from http://www.oecd.org/oecd/pages/home/displaygeneral/0,3380,EN-home-40-1-no-no-no-40,00.html

Pattinson, B., Montagnier, P., & Moussiegt, L. (2000). *Measuring the ICT sector*. Retrieved June 20, 2002, from http://www.oecd.org/pdf/M00002000/M00002651.pdf

Petrin, T., Sicherl, P., Kukar, S., Mesl, M., & Vitez, R. (2000). *Benchmarking Slovenia: An evaluation of Slovenia's competitiveness, strengths and weaknesses*. Ljubljana, Slovenia: Ministry of Economic Affairs.

Pew Research Center (n.d.). *For the people and the press*. Retrieved October 22, 2002, from http://people-press.org

Rao, A., Carr, L. P., Dambolena, I., Kopp, R. J., Martin, J., Rafii, R., & Schlesinger, P. F. (1996). *Total quality management: A cross functional perspective*. New York: Wiley.

Statistical Indicators for Benchmarking the Information Society (2002). Home page. Retrieved October 22, 2002, from http://www.sibis-eu.org/sibis/

Sicherl, P. (2001). *Metodologija*. Ljubljana: Ministrstvo za informacijsko druzbo. Retrieved June 25, 2002, from http://www.gov.si:80/mid/Dokumenti/CasovneDistance/CD_metodologija.pdf

Spendolini, M. J. (1992). *The benchmarking book*. New York: AMACOM.

Statistics Sweden and Eurostat (2001). *Report of the Task Force on Benchmarking in Infra-Annual Economic Statistics to the SPC*. Luxembourg: Eurostat.

Tardugno, A. F., DiPasquale, T. R., & Matthews, R. E. (2000). *IT services: Costs, metrics, benchmarking, and marketing*. Upper Saddle River: Prentice Hall.

United Nations Development Programme (2001). *Human development indicators*. New York: Oxford University Press. Retrieved October 22, 2002, from http://www.undp.org/hdr2001/back.pdf

United Nations Industrial Development Organization (2002). *Industrial development report 2002/2003: Competing through innovation and learning*. Retrieved October 22, 2002, from http://www.unido.org/userfiles/hartmany/12IDR_full_report.pdf

WebWorld (2002). *UNESCO observatory on the information society*. Retrieved October 22, 2002, from http://www.unesco.org/webworld/observatory/about/index.shtml

World Bank Group (2001). *Information and communication technologies (ICT): Sector Strategy Paper*.

FURTHER READING

Benchmarking (n.d.). Retrieved June 22, 2002, from http://www.benchmarking.de

The Benchmarking Exchange (n.d.). Retrieved June 22, 2002, from http://www.benchnet.com

Benchmarking in Europe (n.d.). Retrieved June 22, 2002, from http://www.benchmarking-in-europe.com/index.asp

Benchmarking in Europe Archive (2002). Retrieved June 23, 2002, from http://www.benchmarking-in-europe.com/library/archive_whats_new/index.htm

The Benchmarking Network (n.d.). *The benchmarking* resource guide. Retrieved June 22, 2002, from http://benchmarkingnetwork.com

Best Practices. Global Benchmarking Council (n.d.). Retrieved June 22, 2002, from http://www.globalbenchmarking.com

CAM Benchmarking (2002). *Sectoral benchmarking*. Retrieved June 11, 2002, from http://www.cam-benchmarking.com/nonmem_OUT_Sectoral.asp

Corporate Benchmarking Services (n.d.) Web site. Retrieved June 11, from www.Corporate-Benchmarking.org

Information Systems Management Benchmarking Consortium (n.d.). Web site. Retrieved June 22, 2002, from http://www.ismbc.org

International Data Corporation (2002). *Latin America Internet commerce market model*. Retrieved June 20, 2002, from http://www.idc.com/getdoc.jhtml?sectionId = tables&containerId = LA1142G&pageType = SECTION

International Data Corporation (2002). Web users in Western Europe, 2001–2006. Retrieved June 20, 2002, from http://www.idc.com/getdoc.jhtml?containerId = dg20020704

International Government Benchmarking Association (n.d.). Web site. Retrieved June 22, 2002, from http://www.igba.org

Public Sector Benchmarking Service (n.d.). Web site. Retrieved June 22, 2002, from http://www.benchmarking.gov.uk/default1.asp

Jackson, N., & Lund, H., ed. (2000). *Benchmarking for higher education*. Buckingham, UK: Open University Press.

Kingdom, B. E. A. (1996). *Performance benchmarking for water utilities*. Denver, CO: American Water Woks Association.

Telecommunications International Benchmarking Group (n.d.). Web site. Retrieved June 22, 2002, from http://www.tbig.org

Biometric Authentication

James. L. Wayman, *San Jose State University*

INTRODUCTION

"Biometric authentication" is the automatic identification or identity verification of living humans based on behavioral and physiological characteristics (Miller, 1989). The field is a subset of the broader field of human identification science. Example technologies include, among others, fingerprinting, hand geometry, speaker verification, and iris recognition. At the current level of technology, DNA analysis is a laboratory technique requiring human processing, so it not considered "biometric authentication" under this definition. Some techniques (such as iris recognition) are more physiologically based, some (such as signature recognition) more behaviorally based, but all techniques are influenced by both behavioral and physiological elements.

Biometric authentication is frequently referred to as simply "biometrics," although this term has historically been associated with the statistical analysis of general biological data (*Webster's New World Dictionary,* 1966). The word biometrics is usually treated as singular. In the context of this chapter, biometrics deals with computer recognition of patterns created by human behavior and physiology and is usually associated more with the field of computer engineering than with biology.

APPLICATIONS

The perfect biometric measure would be

- Distinctive: different across users,
- Repeatable: similar across time for each user,
- Accessible: easily displayed to a sensor,
- Acceptable: not objectionable to display by users, and
- Universal: possessed by and observable on all people.

Unfortunately, no biometric measure has all of these attributes: There are great similarities among different individuals, measures change over time, some physical limitations prevent display, "acceptability" is in the mind of the user, and not all people have all characteristics. Practical biometric technologies must compromise on every point. Consequently, the challenge of biometric deployments is to develop robust systems to deal with the vagaries and variations of human beings.

There are two basic applications of biometric systems:

1. To establish that a person is enrolled in a database.
2. To establish that a person is not enrolled in a database.

Immigration systems, amusement parks, and health clubs use biometrics in the first application: to link users to their enrolled identity. Social service, drivers' licensing, and national identification systems use biometrics primarily in the second application: to establish that prospective participants are not already enrolled. Although hybrid systems—using "negative identification" to establish that a user is not already enrolled, then using "positive identification" to recognize enrolled individuals in later encounters—are also possible, they are not common. The largest biometric systems in place, worldwide, are for purely negative identification.

The key to all of these of systems is the "enrollment" process, in which a user presents for the first time one or more biometric measures to be processed and stored by the system. Systems of the first type, called "positive identification systems" do not necessarily require centralized databases but can use distributed storage, such as on individual computers or machine-readable cards. Systems of the second type, called "negative identification systems," require a centralized database or its equivalent.

For positive identification systems using distributed storage, the submitted sample can be compared only to a single template on the storage media. This is called "one-to-one" verification. Positive identification systems using a centralized database may require users to enter an identification number or name (perhaps encoded on a mag-stripe card) to limit the size of the required search to only a portion of the entire database. If the identifying number or name is unique to each enrolled individual, this form of positive verification can also be "one-to-one," even though the centralized database contains many enrolled individuals.

Large-scale negative identification systems generally partition the database using factors such as gender or age so that not all centrally stored templates need be examined to establish that a user is not in the database. Such systems are sometimes loosely called "one-to-N," where *N* represents only a small portion of the enrolled users. In the general case, however, both positive and negative identification systems search one or more submitted samples against many stored templates or models.

All biometric systems can only link a user to an enrolled identity at some incomplete level of certainty. A biometric system can neither verify the truth of the enrolled identity nor establish the link automatically with complete certainty. If required, determining a user's "true" identity is done at the time of enrollment through trusted external documentation, such as a birth certificate or driver's license. When the user is later linked to that enrolled identity through a biometric measure, the veracity of that identity is only as reliable as the original documentation used for enrollment.

All biometric measures may change over time, due to aging of the body, injury, or disease. Therefore, reenrollment may be required. If "true" identity or continuity of identity is required by the system, reenrollment must necessitate presentation of trusted documentation. Not all systems, however, have a requirement to know a user's "true" identity. Biometric measures can be used as identifiers in anonymous and pseudo-anonymous systems.

Although biometric technologies are not commonly used with Internet transactions today, future uses would most likely be in the first application: to establish that a person is enrolled in a database and, therefore, has certain attributes and privileges within it, including access authorization. The argument can be made that biometric measures more closely link the authentication to the human user than passwords, personal identification numbers (PINs), PKI codes, or tokens, which authenticate machines. Consequently, the focus and terminology of this chapter is on applications of the first type, "positive identification."

HISTORY

The science of recognizing people based on physical measurements owes to the French police clerk Alphonse Bertillon, who began his work in the late 1870s (Beavan, 2001; Cole, 2002). The Bertillon system involved at least 11 measurements, such as height, the length and breadth of the head, and length of the ring and middle fingers. Categorization of iris color and pattern was also included in the system. By the 1880s, the Bertillon system was in use in France to identify repeat criminal offenders. Use of the system in the United States for the identification of prisoners began shortly thereafter and continued into the 1920s. Extreme claims of accuracy were made for the system, based on the unsupportable hypothesis that the various measures were statistically independent (Galton, 1890; Galton, 1908).

Although research on fingerprinting by a British colonial magistrate in India, William Herschel, began in the late 1850s, knowledge of the technique did not become known in the Western world until the 1880s (Faulds, 1880; Herschel, 1880) when it was popularized scientifically by Sir Francis Galton (1888) and in literature by Mark Twain (1992/1894). Galton's work also included the identification of persons from profile facial measurements.

By the mid-1920s, fingerprinting had completely replaced the Bertillon system within the U.S. Bureau of Investigation (later to become the Federal Bureau of Investigation). Research on new methods of human identification continued, however, in the scientific world. Handwriting analysis was recognized by 1929 (Osborn, 1929), and retinal scanning was suggested in 1935 (Smith & Goldstein, 1935).

None of these techniques was "automatic," however, so none meets the definition of "biometric authentication" used in this chapter. Automatic techniques require automatic computation. Work in automatic speaker identification can be traced directly to experiments with analog computers done in the 1940s and early 1950s (Chang, Pihl, & Essignmann, 1951). With the digital revolution beginning in the 1950s, a strong tool for human identification through pattern matching became available: the digital computer. Speaker (Atal, 1976; Rosenberg, 1976) and fingerprint (Trauring, 1963a) pattern recognition were among the first applications in digital signal processing. By 1961, a "wide, diverse market" for computer-based fingerprint recognition was identified, with potential applications in "credit systems," "industrial and military security systems," and "personal locks" (Trauring, 1963b). Computerized facial recognition followed (Kanade, 1977). By the mid-1970s, the first operational fingerprint and hand geometry systems (Raphael & Young, 1974) were fielded, and formal biometric system testing had begun (National Bureau of Standards, 1977). Iris recognition systems became available in the mid-1990s (Daugman, 1993). Today there are close to a dozen approaches used in commercially available systems (see Table 1).

SYSTEM DESCRIPTION

Given the variety of applications and technologies, it might seem difficult to draw any generalizations about biometric systems. All such systems, however, have many elements in common. Figure 1 shows a general biometric system consisting of data collection, transmission, signal processing, storage, and decision subsystems (Wayman, 1999). This diagram accounts for both enrollment and operation of positive and negative identification systems.

Table 1 Commercially Available Biometric Technologies

Hand geometry
Finger geometry
Speaker recognition
Iris recognition
Facial imaging
Fingerprinting
Palm printing
Keystroke
Hand vein
Dynamic signature
Verification

Data Collection

Biometric systems begin with the measurement of a behavioral or physiological characteristic. Because biometric data can be one- (speech), two- (fingerprint), or multidimensional (handwriting dynamics), it generally does not involve "images." To simplify the vocabulary used in this chapter, I refer to raw signals simply as "samples."

Key to all systems is the underlying assumption that the measured biometric characteristic is both distinctive among individuals and repeatable over time for the same individual. The problems in measuring and controlling these variations begin in the data collection subsystem.

The user's characteristic must be presented to a sensor. The act of presenting a biometric measure to a sensor introduces a behavioral component to every biometric method because the user must interact with the sensor in the collection environment. The output of the sensor, which is the input sample on which the system is built, is the combination of (a) the biometric measure, (b) the way the measure is presented, and (c) the technical characteristics of the sensor. Both the repeatability and the distinctiveness of the measurement are negatively affected by changes in any of these factors. If a system is to communicate with other systems, the presentation and sensor characteristics must be standardized to ensure that biometric characteristics collected with one system will match those collected on the same individual by another system.

Transmission

Some, but not all, biometric systems collect data at one location but store or process it (or both) at another. Such systems require data transmission over a medium such as the Internet. If a great amount of data is involved, compression may be required before transmission or storage to conserve bandwidth and storage space. Figure 1 shows compression and transmission occurring before the signal processing and image storage. In such cases, the transmitted or stored compressed data must be expanded before further use. The process of compression and expansion generally causes quality loss in the restored signal, with loss increasing with increasing compression ratio. Interestingly, limited compression may actually improve the performance of the pattern recognition software as information loss in the original signal is generally in the less repeatable high-frequency components. The compression

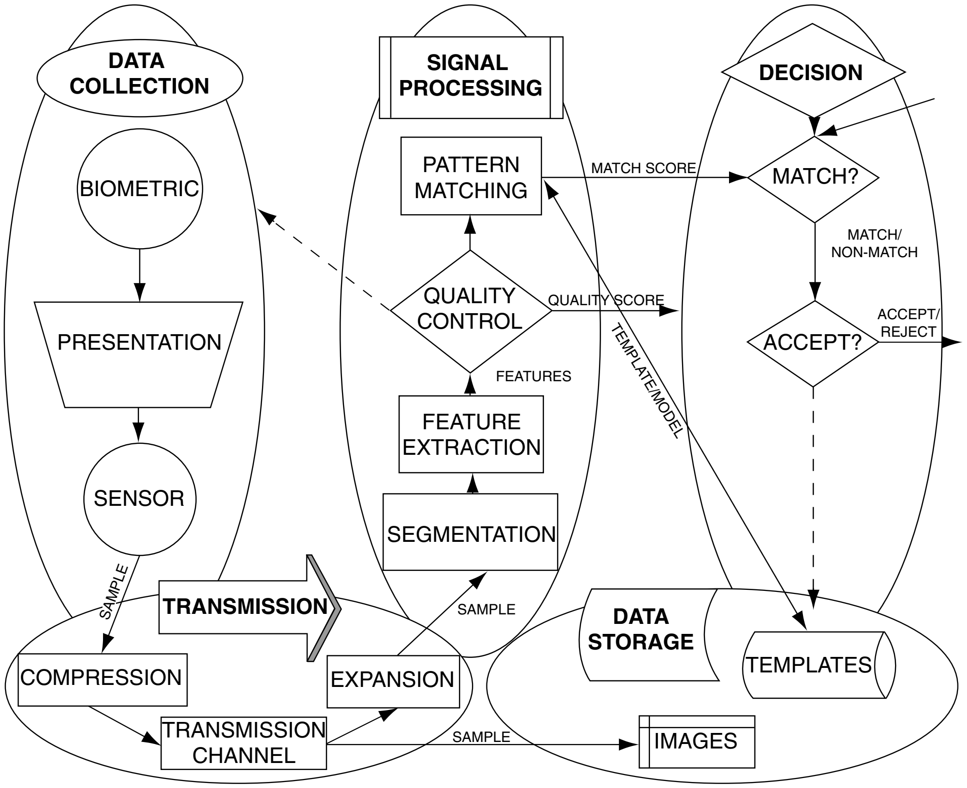


Figure 1: Example biometric system.

technique used will depend on the biometric signal. An interesting area of research is in finding, for a given biometric technique, compression methods that have a minimal impact on the subsequent signal processing activities.

If a system is to allow sharing of data at the signal level with other systems, compression and transmission protocols must be standardized. Standards currently exist for the compression of fingerprint (Wavelet Scalar Quantization, 1993), facial (Information Technology, 1993), and voice (Cox, 1997) data.

Signal Processing

The biometrics signal processing subsystem comprises four modules: segmentation, feature extraction, quality control, and pattern matching. The segmentation module must determine if biometric signals exist in the received data stream (signal detection) and, if so, extract the signals from the surrounding noise. If the segmentation module fails to detect or extract a biometric signal, a "failure-to-acquire" has occurred.

The feature extraction module must process the signal in some way to preserve or enhance the between-individual variation (distinctiveness) while minimizing the within-individual variation (nonrepeatability). The output of this module is numbers, vectors, or distribution parameters that, although called biometric "features," may not have direct physiological or behavioral interpretation. For example, mathematical output from facial recognition systems does not indicate directly the width of the lips or the distances between the eyes and the mouth.

The quality control module must do a statistical "sanity check" on the extracted features to make sure they are not outside population statistical norms. If the sanity check is not successfully passed, the system may be able to alert the user to resubmit the biometric pattern. If the biometric system is ultimately unable to produce an acceptable feature set from a user, a "failure-to-enroll" or a "failure-to-acquire" will be said to have occurred. Failure-to-enroll or acquire may be due to failure of the segmentation algorithm, in which case no feature set will be produced. The quality control module might even affect the decision process, directing the decision subsystem to adopt higher requirements for matching a poor quality input sample.

The pattern matching module compares sample feature sets with enrolled templates or models from the database and produces a numerical "matching score." When both template and features are vectors, the comparison may be as simple as a Euclidean distance. Neural networks might be used instead. Regardless of which pattern matching technique is used, templates or models and features from samples will never match exactly because of the repeatability issues. Consequently, the matching scores determined by the pattern matching module will have to be interpreted by the decision subsystem.

In more advanced systems, such as speaker verification, the enrollment "templates" might be "models" of the signal generation process—very different data structures than the observed features. The pattern matching module determines the consistency of the observed features with the stored generating model. Some pattern matching modules may even direct the adaptive recomputation of features from the input data.

Decision

The decision subsystem is considered independently from the pattern matching module. The pattern matching module might make a simple "match" or "no match" decision based on the output score from the pattern matcher. The decision module might ultimately "accept" or "reject" a user's claim to identity (or nonidentity) based on multiple attempts, multiple measures or a measure-dependent decision criteria. For instance, a "three-try" decision policy will accept a user's identity claim if a match occurs in any of three attempts.

The decision module might also direct operations to the stored database, allowing enrollment templates to be stored, calling up additional templates for comparison in the pattern matching module, or directing a database search.

Because input samples and stored templates will never match exactly, the decision modules will make mistakes—wrongly rejecting a correctly claimed identity of an enrolled user or wrongly accepting the identity claim of an impostor. Thus, there are two types of errors: false rejection and false acceptance. These errors can be traded off against one another to a limited extent: decreasing false rejections at the cost of increased false acceptances and vice versa. In practice, however, inherent within-individual variation (nonrepeatability) limits the extent to which false rejections can be reduced, short of accepting all comparisons. The decision policies regarding "match–no match" and "accept–reject" criteria specific to the operational and security requirements of the system and reflect the ultimate cost of errors of both types.

Because of the inevitability of false rejections, all biometric systems must have "exception handling" mechanisms in place. If exception handling mechanisms are not as strong as the basic biometric security system, vulnerability results. An excessively high rate of false rejections may cause the security level of the exception handling system to be reduced through overload.

The false acceptance rate, on the other hand, measures the percentage of impostors who are able to access the system. The complement of the false acceptance rate is the percentage of impostors who are intercepted. So a 20% false acceptance rate means that 80% of impostors are intercepted. Depending on the application, this may be high enough to serve as a sufficient deterrent to prevent impostors from attempting access through the biometric system. The exception handling mechanism might become a more appealing target for those seeking fraudulent access. Consequently, the security level of a biometric system in a positive identification application may be more dependent on the false rejection rate than the false acceptance rate.

Storage

The remaining subsystem to be considered is that of storage. The processed features or the feature generation model of each user will be stored or "enrolled" in a database for future comparison by the pattern matcher to

Table 2 Biometric Template Sizes

DEVICE	SIZE IN BYTES
Fingerprint	200–1,000
Speaker	100–6,000
Finger geometry	14
Hand geometry	9
Face	100–3,500
Iris	512

incoming feature samples. This enrollment data is called a "template" if it is of the same mathematical type as the processed features and a "model" if it gives a mathematical explanation of the feature generating process. For systems only performing positive identification, the database may be distributed on magnetic strip, optically read, or smart cards that each enrolled user carries. Depending on system policy, no central database for positive identification systems need exist, although a centralized database can be used to detect counterfeit cards or to reissue lost cards without recollecting the biometric measures.

The original biometric measurement, such as a fingerprint pattern, is generally not reconstructable from the stored templates. Furthermore, the templates themselves are created using the proprietary feature extraction algorithms of the system vendor. If it may become necessary to reconstruct the biometric patterns from stored data (to support interoperability with systems from other vendors, for example), raw (although possibly compressed) data storage will be required. The storage of raw data allows changes in the system or system vendor to be made without the need to recollect data from all enrolled users. Table 2 shows some example template sizes for various biometric devices.

PERFORMANCE TESTING

Biometric devices and systems might be tested in many different ways. Types of testing include the following:

- Technical performance;
- Reliability, availability, and maintainability [RAM];
- Vulnerability;
- Security;
- User acceptance;
- Human factors;
- Cost–benefit; and
- Privacy regulation compliance.

Technical performance has been the most common form of testing in the last three decades, and "best practices" have been developed (Mansfield & Wayman, 2001). These tests generally measure failure-to-enroll, failure-to-acquire, false acceptance, false rejection, and throughput rates. Failure-to-enroll rate is determined as the percentage of all persons presenting themselves to the system in "good faith" for enrollment who are unable to do so because of system or human failure. Failure-to-acquire rate is determined as the percentage of "good faith" presentations by all enrolled users that are not acknowledged by the system. The false rejection rate is the percentage of all users whose claim to identity is not accepted by the system. This will include failed enrollments and failed acquisitions, as well as false nonmatches against the user's stored template. The false acceptance rate is the rate at which "zero-effort" impostors making no attempt at emulation are incorrectly matched to a single, randomly chosen false identity. Because false acceptance–rejection and false match–nonmatch rates are generally competing measures, they can be displayed as "decision error trade-off" (DET) curves.

The throughput rate is the number of persons processed by the system per minute and includes both the human–machine interaction time and the computational processing time of the system.

Types of Technical Tests

There are three types of technical tests: technology, scenario, operational (Phillips, Martin, Wilson, & Przybocki, 2000).

Technology Test

The goal of a technology test is to compare competing algorithms from a single technology, such as fingerprinting, against a standardized database collected with a "universal" sensor. There are competitive, government-sponsored technology tests in speaker verification (National Institute of Standards and Technology, 2003), facial recognition (Philips, Grother, Michaels, Blackburn, Tabassi, & Bone, 2003), and fingerprinting (Maio, Maltoni, Wayman, & Jain, 2000).

Scenario Test

Although the goal of technology testing is to assess the algorithm, the goal of scenario testing is to assess the performance of the users as they interact with the complete system in an environment that models a "real-world" application. Each system tested will have its own acquisition sensor and so will receive slightly different data. Scenario testing has been performed by a number of groups, but few results have been published openly (Bouchier, Ahrens, & Wells, 1996; Mansfield, Kelly, Chandler, & Kane, n.d.; Rodriguez, Bouchier, & Ruehie, M., 1993).

Operational Test

The goal of operational testing is to determine the performance of a target population in a specific application environment with a complete biometric system. In general, operational test results will not be repeatable because of unknown and undocumented differences between operational environments. Furthermore, "ground truth" (i.e., who was actually presenting a "good faith" biometric measure) will be difficult to ascertain. Because of the sensitivity of information regarding error rates of operational systems, few results have been reported in the open literature (Wayman, 2000b).

Regardless of the type of test, all biometric authentication techniques require human interaction with a data collection device, either standing alone (as in technology

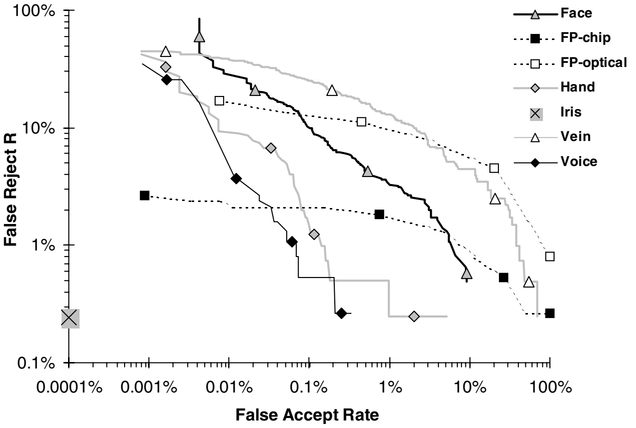


Figure 2: Detection error trade-off curve: Best of three attempts.

testing) or as part of an automatic system (as in scenario and operational testing). Consequently, humans are a key component in all system assessments. Error, failure-to-enroll or acquire, and throughput rates are determined by the human interaction, which in turn depends on the specifics of the collection environment. Therefore, little in general can be said about the performance of biometric systems or, more accurately, about the performance of the humans as they interact with biometrics systems.

The National Physical Lab Tests

A study by the U.K. National Physical Laboratory in 2000 (Mansfield et al., 2001b) looked at eight biometric technologies in a scenario test designed to emulate access control to computers or physical spaces by scientific professionals in a quiet office environment.

The false accept–false reject DET under a "three-tries" decision policy for this test is shown in Figure 2. The false rejection rate includes "failure-to-enroll/acquire" rates in its calculation.

Throughput Rates

The National Physical Laboratory study (Mansfield et al., n.d.) also established transaction times for various biometric devices in an office, physical access control setting, shown as Table 3.

In Table 3, the term "PIN" indicates whether the transaction time included the manual entry of a four-digit identification number by the user. These times referred only to the use of the biometric device and did not include actually accessing a restricted area.

Table 3 Transaction Times in Office Environment

TRANSACTION TIME				
DEVICE	MEAN	MEDIAN	MINIMUM	PIN?
Face	15	14	10	No
Fingerprint—optical	9	8	2	No
Fingerprint—chip	19	15	9	No
Hand	10	8	4	Yes
Iris	12	10	4	Yes
Vein	18	16	11	Yes
Speaker	12	11	10	No

Biometric Forgeries

It has been well known since the 1970s that all biometric devices can be fooled by forgeries (Beardsley, 1972). In a positive identification system, "spoofing" is the use of a forgery of another person's biometric measures. (In a negative identification system, "spoofing" is an attempt to disguise one's own biometric measure.) Forging biometric measures of another person is more difficult than disguising one's own measures, but is possible nonetheless. Several studies (Blackburn et al., 2001; Matsumoto et al., 2002; Thalheim, Krissler, & Ziegler, 2002; van der Putte & Keuning, 2002) discuss ways by which facial, fingerprint and iris biometrics can be forged. Because retinal recognition systems have not been commercially available for several years, there is no research on retina forgery, but it is thought to be a difficult problem. Speaker recognition systems can make forgery difficult by requesting the user to say randomly chosen numbers. The current state of technology does not provide reliable "liveness testing" to ensure that the biometric measure is both from a living person and not a forgery.

EXAMPLE APPLICATIONS

Most successful biometric programs for positive identification have been for physical access control. In this section, I consider two such programs, one at Walt Disney World and the other the U.S. INSPASS program.

Disney World

The Walt Disney Corporation needed to link guests to season passes without inconveniencing passholders on visits to their four Orlando parks (Disney World, Epcot, Animal Kingdom, and MGM Studio Tour) and two water parks (Blizzard Beach and Typhoon Lagoon) (Levin, 2002). The challenge was to create a verification component with minimum impact on existing systems and procedures. Alternatives to biometrics existed for Disney. They considered putting photos on the passes, then checking photos against passholders. This approach required human inspectors at all the entrances. Disney could have put names on passes at the point of sale, then checked passes against photo identification presented by the pass holders at the entrance to the parks, but this would have required disclosure of the passholders' identities.

Automatic, anonymous authentication of the passholders might have been accomplished by requiring passholders to enter an identifying number of some sort upon entrance to the parks. That number would be chosen at the time of sale or first use of the pass and could be chosen by the passholder. The concerns with this approach are twofold: forgotten numbers and the increased guest processing time. Disney seeks to make parks as accessible as possible, and use of key pads could create accessibility problems. Disney ultimately decided to use biometric authentication, taking the shapes of right index and middle finger on all passholders. Finger shape only is recorded.

Fingerprints are not imaged. The template is established when the holder places the fingers in the imaging device upon first use of the pass at the park entrance. This first user becomes the "authorized holder" of the pass. The template is stored centrally in the Disney access control system and is linked to the pass number encoded on a magnetic stripe on the pass. Upon subsequent uses, the card number allows the recall of the stored finger geometry, which is then compared with that presented by the guest to determine if the guest matches the original holder of the pass. A failure of the system to authenticate the guest is assumed to be a system error and a guest relations officer is on hand to resolve the issue quickly. Disney considers the system to be quite successful.

INSPASS

Frequent travelers to the United States can bypass immigration lines if they hold a card issued by the Immigration and Naturalization Service's (INS) Passenger Accelerated Service System (INSPASS). The system is currently in place at Kennedy, Newark, Los Angeles, Miami, San Francisco, Detroit, and Washington Dulles airports, as well as in Vancouver, and Toronto airports in Canada. Users give measurements of their hand geometry at the INSPASS kiosk to verify that they are the correct holder of the card. The passport number on the card can then be trusted by the INS to be that of the holder and the border crossing event is automatically recorded.

At the nine airports with INSPASS, there are more than 21 kiosks, and more than 20,000 automated immigration inspections were conducted on average each month in 2000. About 72% of these inspections were for U.S. citizens, but any citizen of any country on the U.S. "visa waver program" can apply for and receive an INSPASS card. In 2000, there were more than 45,000 active enrollees, and each traveler used INSPASS about four times annually, on average. As of this writing, INSPASS is temporarily on hiatus at some airports, pending the creation of a thorough business plan for the system by the INS Office of Inspections.

Use of INSPASS is entirely voluntary. Applicants attest that they have no drug or smuggling convictions and supply fingerprints, along with their hand geometry, at the time of enrollment. A passport is also required as proof of "true" identity. The INS considers the INSPASS holders to be low-risk travelers already well-known to the system and therefore affords them the special privilege of entering the United States without a face-to-face meeting with an immigration officer. Those not using INSPASS have passports inspected by an immigration officer, whose duty it is to ferret out those entering illegally or on forged or stolen passports. Immigration officers are also charged with prescreening all arrivals for customs, agriculture, and State Department–related issues. Returning U.S. citizens not holding INSPASS cards must answer questions such as length of trip, travel destinations, and purpose of travel. Non-U.S. citizens arriving without an INSPASS must answer an even more extensive series of questions about length of stay, intent of visit, and destinations within the United States. In effect, INSPASS substitutes the possession of card for the passport and substitutes presentation of the hand geometry for the usual round of border crossing questions. Analyses of the error rates for this system are given in Wayman (2000b).

BIOMETRICS AND PRIVACY

The concept "privacy" is highly culturally dependent. Legal definitions vary from country to country and, in the United States, even from state to state (Alderman & Kennedy, 1995). A classic definition is the intrinsic "right to be let alone," (Warren & Brandeis, 1980), but more modern definitions include informational privacy: the right of individuals "to determine for themselves when, how and to what extent information about them is communicated to others" (Westin, 1967). The U.S. Supreme Court has recognized both intrinsic (*Griswald v. Connecticut,* 1967) and informational (*Whalen v. Roe,* 1977) privacy. Both types of privacy can be affected positively or negatively by biometric technology.

Intrinsic (or Physical) Privacy

Some people see the use of biometric devices as an intrusion on intrinsic privacy. Touching a publicly used biometric device, such as a fingerprint or hand geometry reader, may seem physically intrusive, even though there is no evidence that disease can spread any more easily by these devices than by door handles. People may also object to being asked to look into a retinal scanner or to stand still while giving an iris or facial image. Not all biometric methods require physical contact. A biometric application that replaced the use of a keypad with the imaging of an iris, for instance, might be seen as enhancing of physical privacy.

If biometrics are used to limit access to private spaces, then biometrics can be more enhancing to intrinsic privacy than other forms of access control, such as keys, which are not as closely linked to the holder.

There are people who object to use of biometrics on religious grounds: Some Muslim women object to displaying their face to a camera, and some Christians object to hand biometrics as "the sign of the beast" (Revelations, 13:16–18). It has been noted (Seildarz, 1998), however, that biometrics are the marks given an individual by God.

It can be argued (Baker, 2000; Locke, 1690) that a physical body is not identical to the person that inhabits it. Whereas PINs and passwords identify persons, biometrics identifies the body. Some people are uncomfortable with biometrics because of this connection to the physical level of human identity and the possibility of nonconsentual collection of some biometric measures. Biometric measures could allow linking of the various "persons" or psychological identities that each of us choose to manifest in our separate dealings within our social structures. Biometrics, if universally collected without adequate controls, could aid in linking my employment records to my health history and church membership. This leads us to the concept of "informational privacy."

Informational Privacy

With notably minor qualifications, biometric features contain no personal information whatsoever about the user. This includes no information about health status, age, nationality, ethnicity, or gender. Consequently, this also limits the power of biometrics to prevent underage access to pornography on the Internet (Woodward, 2000) or to detect voting registration by noncitizens (Wayman 2000a).

No single biometric measure has been demonstrated to be distinctive or repeatable enough to allow the selection of a single person out of a very large database. (All automatic large-scale civilian fingerprint systems, for instance, require images from multiple fingers of each user for unique identification. Law enforcement automatic fingerprint identification systems (AFIS) require human intervention to identify an individual from "latent" fingerprint images left at crime scenes.) When aggregated with other data, however, such as name, telephone area code, or other even weakly identifying attributes, biometric measures can lead to unique identification within a large population. For this reason, databases of biometric information must be treated as personally identifiable information and protected accordingly.

Biometrics can be directly used to enhance informational privacy. Use to control access to databases containing personal and personally identifiable data can enhance informational privacy. The use of biometric measures, in place of name or social security number, to ensure that personal data is anonymous, is privacy enhancing.

Biometric measures are private but not secret. The U.S. Supreme Court has ruled that "Like a man's facial characteristics, or handwriting, his voice is repeatedly produced for others to hear" (*U.S. v. Dionisio,* 1973). Therefore, the court concluded, no can reasonably expect that either his or her voice or face "will be a mystery to the world." The theft of biometric measures through reproduction, theft of identity at biometric enrollment, or theft of biometrically protected identity by data substitution through reenrollment could all have grave repercussions for both intrinsic and informational privacy.

STANDARDS

Biometric methods, such as fingerprinting, face, and speaker recognition, were developed independently, by different academic traditions and for different applications. It has only been within the last decade that the commonality as automatic human identification methods has even been recognized, so it should come as no surprise that common standards have been slow to develop.

Furthermore, there has been little need for interoperability among these systems. In fact, the lack of interoperability within and across technologies has been touted as a privacy-preserving asset of biometric systems (Woodward, 1999). Consequently, there has been no motivation for the tedious standards development process required to promote interoperability. A notable exception is in automatic fingerprint identification systems (AFIS), for which law enforcement has long needed the capability of interjurisdictional fingerprint exchanges. In this case, some American National Standards Institute (ANSI), National Institute of Standards and Technology (NIST), and Criminal Justice Information Services (CJIS) recognized standards do exist. These have been accepted as de facto international standards.

In 2002, the International Standards Organization/International Electrotechnical Commission Joint Technical Committee 1 formed Standing Committee 37 to create international standards for biometrics. Early work by SC37 has focused on "harmonized vocabulary and definitions," "technical interfaces," "data interchange formats," "application profiles," "testing and reporting," and "cross jurisdictional and societal aspects."

Fingerprint Standards

The CJIS (1999) report Interim IAFIS Fingerprint Image Quality Specifications for Scanners specifies technical requirements (signal-to-noise ratio, gray scale resolution, etc.) for the scanning of fingerprint images for transmission to the FBI. This standard, commonly known as "Appendix F/G," was developed for "flat bed scanners" used for the digital imaging of inked fingerprint cards. It has been applied, with difficulty, as a standard for fingerprint sensors used in large-scale biometric fingerprint systems. Fingerprint devices not used for large-scale searches, such as those for computer or facilities access control, commonly do not meet "Appendix F/G" technical standards.

Fingerprint images, even when compressed, are around 15 kBytes in size. As shown in Table 2, templates extracted using proprietary algorithms are much smaller. The requirement to store fingerprint information on interoperable ID cards has lead to some efforts at template standardization the American Association of Motor Vehicle Administrators (AAMVA, 2000), motivated by the need for cross-jurisdictional exchange of fingerprint data for driver identification, has published a standard for drivers licenses and identification cards, which includes a section on fingerprint minutiae extraction and storage. Although this standard has never been tested or implemented, it is hoped that it will provide a possible solution to fingerprint system interoperability.

Facial Image Standards

Compression standards for facial imaging have already been mentioned (Information Technology, 1993). NIST has a document "Best Practice Recommendations for the Capture of Mugshots" (NIST, 1993), that has application to facial recognition when "mug shot" type photos are used.

BioAPI

Beyond the issue of interoperability standards is that of software protocol conventions. In the past, each vendor has used its own platform-specific software to support its own data collection, feature extraction, and storage and matching operations. This has caused major headaches for system integrators who try to use biometric devices in larger or more general access control and information retrieval systems. The integrators have been forced to learn and handle the idiosyncratic software of each biometric vendor. During the last 5 years, under the sponsorship of both the U.S. Department of Defense (Human Authentication-Application Programming Interface Steering Group, 1998) and NIST, common software standards have been emerging. This effort is currently known as the "Biometric Applications Programming Interface" (BioAPI Consortium, 2001). This standard specifies exactly how information will be passed back and forth between the larger system and the biometric subsystems. It will allow system integrators to establish one set of software function calls to handle any biometric device within the system.

CBEFF

Closely related to the BioAPI effort is the Common Biometric Exchange File Format (CBEFF) working group sponsored by NIST and the National Security Agency (NIST, 2001). This group has developed a standard "packaging" protocol for transferring biometric templates and samples between applications. Recently, this work is being extended by the Organization for the Advancement of Structured Information Standards (http:// www.oasis-open.org) to an extensible markup language (XML) format for biometric transmission over the Internet.

ANSI X9.84

The ANSI X9.84–2001 standard for Biometric Information Management and Security (ANSI, 2001) is a comprehensive document describing proper architectures and procedures for the financial industry when using biometrics for secure remote electronic access or local physical access control. The standard discusses security requirements for the collection, distribution, and processing of biometric data; uses of biometric technology; techniques for secure transmission of biometric data; and requirements for data privacy.

POTENTIAL INTERNET APPLICATIONS

There have been few direct applications of biometrics to the Internet. Biometrics can and are being used to control local network and PC log-on, however. Many commercial devices are available for this application, including fingerprint, keystroke, speaker, and iris recognition. Some of these products have in the past been marketed at office-supply chain stores. Users enroll biometric data on their local network or own PC. At log-on, submitted biometric data is compared locally with that locally stored. This model can be extended to include biometric authorization for access to computer files and services. No biometric data ever leaves the PC or local network.

How might biometrics be used directly over the Internet? One model requires users to supply biometric data over the internet when registering at a Web site. Of course, such a system has no way of determining the true identity of the user or even whether the biometric data supplied really belongs to the registrant. But once enrolled, the Web site could determine whether the user's browser had access to the same biometric measures given at enrollment. This might indicate continuity of the person at the browser or across browsers.

Recognizing the weakness of such a system, one speaker verification company established a business model of "out-of-band" registration over the telephone. Those registering would speak to a human operator, who would collect identification information, then switch the registrant to a computer for collecting speech samples. When registered users wished to access a controlled Web site, they would telephone the speaker verification company, verify their voice, and be issued an alphanumeric "volatile pass code" that would allow access to the requested site if promptly submitted over the internet.

A third model would be a centralized "biometric certification authority" (BCA), operated either by government or a commercial entity. Users would enroll a biometric measure in person, or possibly through an "out-of-band" mechanism such as the mail. Depending on the purpose of the system, users might be required to present proof of identity, credit card information, or the like. After registration, users wishing to access biometrically controlled Web sites would submit their biometric measures over the Internet to the BCA, which would verify identity to the Web site being accessed. Fear of centralized biometric databases in the hands of either business or government has inhibited implementation of this model.

Given the nonrevocability of biometric information, its "private but not secret" status, and the need for in-person registration when applied to user identification, it is not clear that biometrics has a secure place in Internet communication or commerce.

Commonsense Rules for Use of Biometrics

From what has been discussed thus far, we can develop some commonsense rules for the use of biometrics for access control and similar positive identification applications.

- Never participate in a biometric system that allows either remote enrollment or reenrollment. Such systems have no way of connecting a user with the enrolled biometric record, so the purpose of using biometrics is lost.
- Biometric measures can reveal people's identity only if they are linked at enrollment to their name, social security number, or other closely identifying information. Without that linkage, biometric measures are anonymous.
- Remember that biometric measures cannot be reissued if stolen or sold. Do not enroll in a nonanonymous system unless you have complete trust in the system administration.
- All biometric access control systems must have "exception handling" mechanisms for those that either cannot enroll or cannot reliably use the system. If you are uncomfortable with enrolling in a biometric system for positive identification, insist on routinely using the "exception handling" mechanism instead.
- The safest biometric systems are those in which each user controls his or her own template.
- Because biometric measures are not perfectly repeatable, are not completely distinctive, and require specialized data collection hardware, biometric systems are not useful for tracking people. Anyone who really wants to physically track you will use your credit card, phone records, or cell phone emanations instead. Anyone wanting to track your Internet transactions will do so with cookies or Web logs.

CONCLUSION

Automated methods for human identification have a history predating the digital computer age. For decades, mass adoption of biometric technologies has appeared to be just a few years away (Raphael & Young, 1974; The Right Biometric, 1989), yet even today, difficulties remain in establishing a strong business case, in motivating consumer demand, and in creating a single system usable by all sizes and shapes of persons. Nonetheless, the biometric industry has grown at a steady pace as consumers, industry, and government have found appropriate applications for these technologies. Testing and standards continue to develop, and the privacy implications continue to be debated. Only time will tell if biometric technologies will extend to the Internet on a mass scale.

CROSS REFERENCES

See *Digital Identity; Internet Security Standards; Passwords; Privacy Law.*

GLOSSARY

Biometrics The automatic identification or identity verification of living humans based on behavioral and physiological characteristics.

Biometric measures Numerical values derived from a submitted physiological or behavioral sample.

Decision A determination of probable validity of a user's claim to identity or nonidentity in the system.

Enrollment Presenting oneself to a biometric system for the first time, creating an identity within the system, and submitting biometric samples for the creation of biometric measures to be stored with that identity.

Failure-to-acquire rate The percentage of transactions for which the system cannot obtain a usable biometric sample.

Failure-to-enroll rate The percentage of a population that is unable to give repeatable biometric measures on any particular device.

False accept rate (FAR) The expected proportion of transactions with wrongful claims of identity (in a positive ID system) or nonidentity (in a negative ID system) that are incorrectly confirmed. In negative identification systems, the FAR may include the failure-to-acquire rate.

False match rate (FMR) The expected probability that an acquired sample will be falsely declared to match to a single, randomly selected, nonself template or model.

False non-match rate (FNMR) The expected probability that an acquired sample will be falsely declared not to match a template or model of that measure from the same user.

False reject rate (FRR) The expected proportion of transactions with truthful claims of identity (in a positive ID system) or nonidentity (in a negative ID system) that are incorrectly denied. In positive identification systems, the FRR will include the failure-to-enroll and the failure-to-acquire rates.

Features A mathematical representation of the information extracted from the presented sample by the signal processing subsystem that will be used to construct or compare against enrolment templates (e.g., minutiae coordinates, principal component coefficients and iris codes are features).

Genuine claim of identity A truthful positive claim by a user about identity in the system. The user truthfully claims to be him- or herself, leading to a comparison of a sample with a truly matching template.

Impostor claim of identity A false positive claim by a user about identity in the system. The user falsely claims to be someone else enrolled in the system, leading to the comparison of a sample with a nonmatching template.

Identifier An identity pointer, such as a biometric measure, PIN (personal identification number), or name.

Identify To connect a person to an identifier in the database.

Identity An information record about a person, perhaps including attributes or authorizations or other pointers, such as names or identifying numbers.

Identification The process of identifying, perhaps requiring a search of the entire database of identifiers; consequently, the process of connecting a person to a pointer to an information record.

Matching score A measure of similarity or dissimilarity between a presented sample and a stored template.

Models Mathematical representation of the generating process for biometric measures.

Negative claim of identity The claim (either implicitly or explicitly) of a user not to be known to or enrolled in the system. Enrollment in social service systems open only to those not already enrolled is an example.

Positive claim of identity The claim (either explicitly or implicitly) of a user to be enrolled in or known to the system. An explicit claim might be accompanied by a claimed identifier in the form of a name or identification number. Common access control systems are an example.

Sample A biometric signal presented by the user and captured by the data collection subsystem (e.g. fingerprint, face samples, and iris images are samples).

Template A user's stored reference measure based on features extracted from samples.

Transaction An attempt by a user to validate a claim of identity or nonidentity by consecutively submitting one or more samples, as allowed by the system policy.

Verification Proving as truthful a user's positive claim to an identity in the system by comparing a submitted biometric sample to a template or model stored at enrollment.

REFERENCES

Alderman, E., & Kennedy, C. (1995). *The Right to Privacy.* New York: Vintage.

American Association of Motor Vehicle Administrators (2000). National standard for driver's license/identification card (AAMVA 2000-06-30). Retrieved March 15, 2003, from http://www.aamva.org/Documents/stdAAMVADLIDStandrd000630.pdf

American National Standards Institute (2001). Biometric information management and security, X9.84–2001.

Atal, B. S. (1976). Automatic recognition of speakers from their voices. *Proceedings of the IEEE, 64,* 460–487.

Baker, L. (2000). *Persons and bodies: A constitution view.* Cambridge, England: Cambridge University Press.

Beavan, C. *Fingerprints.* New York: Hyperion.

Beardsley, C. (1972, January). Is your computer insecure? *IEEE Spectrum,* 67–78.

BioAPI Consortium (2001, March 16). *BioAPI specifications, Version 1.1.* Retrieved March 15, 2003, from www.bioapi.org

Blackburn, D., Bone, M., Grother, P., & Phillips, J. (2001, February). Facial recognition vendor test 2000: Evaluation report. U.S. Department of Defense. Retrieved March 15, 2003, from http://www.frvt.org/DLs/FRVT_2000.pdf

Bouchier, F., Ahrens, J., & Wells, G. (1996, April). Laboratory evaluation of the IriScan prototype biometric identifier. Sandia National Laboratories, Report SAND96–1033. Retrieved March 15, 2003, from http://infoserve.library.sandia.gov/sand_doc/1996/961033.pdf

Chang, S. H., Pihl, G. E., & Essignmann, M. W. (1951, February). Representations of speech sounds and some of their statistical properties. *Proceedings of the IRE,* 147–153.

Cole, S. *Suspect identities.* Cambridge, MA: Harvard University Press.

Cox, R. (1997). Three new speech coders for the ITU cover a range of applications. *Communications, 35*(9), 40–47.

Criminal Justice Information Services (1999, January 29). Appendix G: Interim IAFIS image quality specifications for scanners. In *Electronic Fingerprint Transmission Specification,* CJIS-RS-0010(v7). Retrieved March 15, 2003, from http://www.fbi.gov/hq/cjisd/iafis/efts70/cover.htm

Daugman, J. (1993). High confidence visual recognition of persons by a test of statistical independence. *Transactions on Pattern Analysis and Machine Intelligence, 15,* 1148–1161.

Faulds, H. (1880). On the skin furrows of the hand. *Nature, 22,* 605.

Galton, F. (1888). On personal identification and description. *Nature. 38,* 173–177, 201–202.

Galton, F. (1890). Kinship and correlation. *North American Review, 150,* 419–431.

Galton, F. (1908). *Memories of my life.* London: Methuen.

Griswald v. Connecticut, 381 U.S. 479 (1965).

Human Authentication–Application Programming Interface Steering Group (1998, June 30). Meeting notes. Retrieved June 1, 2002, from www.biometrics.org/REPORTS/HAAPI20/sg2c95.zip

Herschel, W. J. (1880). Skin furrows of the hand. *Nature, 23,* 76.

Information technology—Digital compression and coding of continuous-tone still images: Requirements and guidelines (1993). CCITT recommendation T.81, ISO/IEC—10918. Retrieved March 15, 2003, from http://www.w3.org/Graphics/JPEG/itu-t81.pdf

Kanade, T. (1977). *Computer recognition of human faces.* Stuttgart: Birkhauser.

King, S., Harrelson, H., & Tran, G. (2002, February 15). *Testing iris and face recognition in a personnel identification application.* Biometric Consortium. Retrieved March 15, 2003, from www.itl.nist.gov/div895/isis/bc2001/FINAL_BCFEB02/FINAL_1_Final%20Steve%20King.pdf

Levin, G. (2002, February). Real world, most demanding biometric system usage. *Proceedings of the Biometrics Consortium, 2001/02.* Retrieved March 15, 2003, from http://www.itl.nist.gov/div895/isis/bc2001/FINAL_BCFEB02/FINAL_4_Final%20Gordon%20Levin%20Brief.pdf

Locke, J. (1690). *An essay concerning human understanding* (book 2, chapter 27). Retrieved March 15, 2003, from http://www.ilt.columbia.edu/publications/locke_understanding.html

Maio, D., Maltoni, D., Wayman, J., & Jain, A. (2000, September). FVC2000: Fingerprint verification competition 2000. *Proceedings of the 15th International Conference on Pattern Recognition,* Barcelona. Retrieved March 15, 2003, from http://www.csr.unibo.it/fvc2000/download.asp

Mansfield, A., Kelly, G., Chandler, D., & Kane, J. (2001, March). *Biometric product testing final report.* National Physical Laboratory report for Communications Electronic Security Group and the Biometrics Working Group. Retrieved March 15, 2003, from http://www.cesg.gov.uk/technology/biometrics/media/Biometric Test Report pt1.pdf

Mansfield, M. J., & Wayman, J. L. (2002, February). Best practices for testing and reporting biometric device performance. Issue 2.0, U.K. Biometrics Working Group. Retrieved March 15, 2003, from http://www.cesg.gov.uk/technology/biometrics/media/Best Practice.pdf

Matsumoto, T., Matsumoto, H., Yamada, K., & Hoshino, S. (2002, January). Impact of artificial "gummy" fingers on fingerprint systems. *Proceedings of SPIE–The International Society of Optical Engineering,* 4677. Retrieved March 15, 2003, from http://cryptome.org/gummy.htm

Miller, B. L. (1989). Everything you need to know about automated biometric identification. *Biometric Industry Directory.* Washington, DC: Warfel and Miller.

National Bureau of Standards (1977). Guidelines on evaluation of techniques for automated personal identification. Federal Information Processing Standard Publication 48. Washington, DC: National Bureau of Standards.

National Institute of Standards and Technology (2003). The NIST year 2003 speaker recognition evaluation plan. Retrieved March 15, 2003, from http://www.nist.gov/speech/tests/spk/2003/doc/2003-spkrecevalplan-v2.2.pdf

National Institute of Standards and Technology Image Group (1993). Best practice recommendations for capturing mugshots and facial images, version 2. Retrieved March 15, 2003, from http://www.itl.nist.gov/iad/vip/face/bpr_mug3.html

National Institute of Standards and Technology (2001, January). The common biometric exchange file format, NISTIR 6529. Retrieved March 15, 2003, from http://www.itl.nist.gov/div895/isis/bc/cbeff/CBEFF010301web.PDF

Osborn, S. (1929). *Questioned documents*. Chicago: Nelson-Hall.

Phillips, P. J., Martin, A., Wilson, C. L., & Przybocki, M. (2000). An introduction to evaluating biometric systems. *Computer, 33,* 56–63. Retrieved March 15, 2003, from http://www.frvt.org/DLs/FERET7.pdf

Phillips, P. J., Grother, P., Micheals, R. J., Blackburn, D. M., Tabassi, E., & Bone, M. (2003). Facial recognition vendor test 2002, National Institute of Standards and Technology, NISTIR 6965, March 2003. Retrieved March 15, 2003, from http:/frvt.org

Raphael, D. E. & Young, J. R. (1974). Automated personal identification. Menlo Park, CA: Stanford Research Institute.

Rodriguez, J. R., Bouchier, F., & Ruehie, M. (1993). Performance evaluation of biometric identification devices. Albuquerque, NM: Sandia National Laboratory Report SAND93–1930.

Rosenberg, A. (1976). Automatic speaker verification: A review. *Proceedings of the IEEE, 64,* 475–487.

Seildarz, J. (1998, April 6). Letter to the editor. *Philadelphia Inquirer.*

Simon, C., & Goldstein, I. (1935). A new scientific method of identification. *New York State Journal of Medicine, 35,* 901–906.

Thalheim, L., Krissler, J., & Ziegler, P. (2002, May). Biometric access protection devices and their programs put to the test. *C'T Magazine 11,* 114. Retrieved from March 15, 2003, http://www.heise.de/ct/english/02/11/114

The right biometric for the right application. But when? *Personal Identification News, 5,* 6.

Trauring, M. (1963a). On the automatic comparison of finger ridge patterns. *Nature, 197,* 938–940.

Trauring, M. (1963b, April). Automatic comparison of finger ridge patterns (Report No. 190). Malibu, CA: Hughes Research Laboratories. (Original work published 1961.)

Twain, M. (1990). *Puddin'head Wilson and other tales: Those extraordinary twins and the man that corrupted Hadleyburg.* Oxford: Oxford University Press. (Original work published 1894.)

U.S. v. Dionisio, 410 U.S. 1 (1973).

van der Putte, T., & Keuning, J. (2000, September). *Biometrical fingerprint recognition: Don't get your fingers burned.* Smart Card Research and Advanced Applications, IFIP TC8/WG.8., Fourth Working Group Conference on Smart Card Research and Advanced Applications, pp. 289–303. London: Kluwer Academic.

Warren, S., & Brandeis, L. The right of privacy. (1890). *Harvard Law Review 4.* Retrieved March 15, 2003, from http://www.lawrence.edu/fac/boardmaw/Privacy_brand_warr2.html

Wavelet scalar quantization (WSQ) gray-scale fingerprint image compression specification (1993, February 16). Criminal Justice Information Services, Federal Bureau of Investigation, IAFIS-IC-0110v2.

Wayman, J. M. (1999). Technical testing and evaluation of biometric identification devices. In A. Jain, R. Bolle, & S. Pankanti, *Biometrics: Personal security in networked society.* Boston: Kluwer Acadenuc.

Wayman, J. L. (2000a). Biometric identification technologies in election processes. In J. L. Wayman (Ed.), *U.S. National Biometric Test Center collected works: 1997–2000.* San Jose, CA: San Jose State University. Retrieved March 15, 2003, from http://www.engr.sjsu.edu/biometrrics/nbtccw.pdf

Wayman, J. L. (2000b). Evaluation of the INSPASS hand geometry data. In J. L. Wayman (Ed.), *U.S. National Biometric Test Center collected works: 1997–2000.* San Jose, CA: San Jose State University. Retrieved March 15, 2003, from http://www.engr.sjsu.edu/biometrrics/nbtccw.pdf

Webster's new world dictionary of the American language (1966). New York: World.

Westin, A. (1967). *Privacy and freedom.* Boston: Atheneum.

Whalen v. Roe, 429 U.S. 589 (1977).

Woodward, J. (1999). Biometrics: Privacy's foe or privacy's friend? In A. Jain, R. Bolle, & S. Pankanti (Eds.), *Biometrics: Personal security in networked society*. Boston: Kluwer Academic.

Woodward, J. D., Jr. (2000, June 9). Age verification technologies (testimony). Hearing before the Commission on Online Child Protection (COPA). Retrieved March 15, 2003, from http://www.copacommission.org/meetings/hearing1/woodward.test.pdf

Bluetooth™—A Wireless Personal Area Network

Brent A. Miller, *IBM Corporation*

INTRODUCTION

Launched in May 1998, Bluetooth™ wireless technology rapidly has become one of the most well-known means of communication in the information technology industry. The unusual name *Bluetooth* itself has garnered much attention (I discuss the origins of this name later), but the main reason for the focus that the technology receives from so many companies and individuals is the new capabilities that it brings to mobile computing and communication.

This chapter discusses many facets of Bluetooth wireless technology—its origins, the associated Bluetooth Special Interest Group, its applications, especially in personal-area networks, how it works, and how it relates to other wireless technologies. I also present numerous references where more information can be found about this exciting new way to form wireless personal area networks (WPANs) that allow mobile devices to communicate with each other.

BLUETOOTH WIRELESS TECHNOLOGY

Bluetooth wireless technology uses radio frequency (RF) to accomplish wireless communication. It operates in the 2.4-GHz frequency spectrum; the use of this frequency range allows Bluetooth devices to be used virtually worldwide without requiring a license for operation. Bluetooth communication is intended to operate over short distances (up to approximately 100 m, although the nominal range used by most Bluetooth devices is about 10 m). Restricting communication to short ranges allows for low-power operation, so Bluetooth technology is particularly well suited for use with battery-powered personal devices that can be used to form a WPAN. Both voice and data can be carried over Bluetooth communication links, making the technology suitable for connecting both computing and communication devices, such as mobile phones, personal digital assistants (PDAs), pagers, and notebook computers. Table 1 summarizes these key attributes of Bluetooth wireless technology.

Bluetooth wireless technology originally was designed for cable replacement applications, intended to remove the need for a cable between any two devices to allow them to communicate. For example, a cable might be used to connect two computers to transfer files, to connect a PDA cradle to a computer to synchronize data, or to connect a headset to a telephone for hands-free voice calls. This sort of wired operation can often be cumbersome, because the cables used are frequently special-purpose wires intended to connect two specific devices; hence, they are likely to have special connectors that make them unsuitable for general-purpose use. This can lead to “cable clutter”—the need for many cables to interconnect various devices. Mobile users may find this especially burdensome because they need to carry their device cables with them to connect the devices when they are away from home, and even with a large collection of cables, it is unlikely that all of the devices can be plugged together. Nonmobile environments, too, can suffer from cable clutter. In a home or office, wires used to connect, say, computer peripherals or stereo speakers limit the placement of these items, and the cables themselves become obstacles.

Bluetooth technology attempts to solve the problem of cable clutter by defining a standard communication mechanism that can allow many devices to communicate with each other without wires. The next section explores the genesis and evolution of Bluetooth wireless communication.

Origins

The genesis of Bluetooth wireless technology generally is credited to Ericsson, where engineers were searching for a method to enable wireless headsets for mobile telephones. Realizing that such a short-range RF technology could have wider applications, and further realizing that its likelihood for success would be greater as an industry standard rather than a proprietary technology, Ericsson approached other major telephone, mobile computer, and electronics companies about forming an industry group to specify and standardize a general-purpose, short-range, low-power form of wireless communication. This small group became the Bluetooth Special Interest Group (SIG), discussed later.

Of special interest to many is the name “Bluetooth.” Such a name for an industry initiative is unusual. A two-part newsletter article (Kardach, 2001) offers a full explanation of the name’s origin; the salient points follow here. Some of the first technologists involved in early

Table 1 Key Bluetooth Technology Characteristics

CHARACTERISTIC	BLUETOOTH TECHNOLOGY ATTRIBUTES
Medium	Radio frequency (RF) in the 2.4-GHz globally unlicensed spectrum
Range	Nominally 10 m; optionally up to 100 m
Power	Low-power operation, suitable for battery-powered portable devices
Packet types	Voice and data
Types of applications	Cable replacement, wireless personal-area networks
Example applications	Network access, wireless headsets, wireless data transfer, cordless telephony, retail and m-commerce, travel and mobility, and many other applications

discussions about a short-range wireless technology were history buffs, and the discussion at some point turned to Scandinavian history. A key figure in Scandinavian history is 10th-century Danish king Harald Blåtand, who is credited with uniting parts of Scandinavia. It is said that a loose translation of his surname to English produces "blue tooth." Those involved in early discussions of this technology recognized that it could unite the telecommunications and information technology (IT) industries, and hence they referred to it as "Bluetooth," after King Harald. At that time, Bluetooth was considered a temporary "code name" for the project. When the time came to develop an official name for the technology and its associated special interest group, the name Bluetooth was chosen after considering several alternatives. Today this is the trademarked name of the technology and the incorporated entity (the Bluetooth SIG, discussed next) that manages it. In fact, the SIG publishes rules and guidelines (Bluetooth SIG, 2002a) for using the term.

The Bluetooth Special Interest Group

Formed in early 1998 and announced in May of that year, the Bluetooth SIG originally was a rather loosely knit group of five companies: Ericsson, Intel, IBM, Nokia and Toshiba. These companies established themselves as Bluetooth SIG promoter members and formed the core of the SIG. Other companies were invited to join as adopter members, and the SIG's membership grew rapidly. The promoter companies, along with a few invited experts, developed the original versions of the Bluetooth specification (detailed later).

In December 1999, four additional companies—3Com, Lucent, Microsoft, and Motorola—were invited to join the group of promoter companies (later Lucent's promoter membership was transferred to its spin-off company, Agere Systems). By this time, the SIG's membership had grown to more than 2,000 companies. In addition to promoters and adopters, a third membership tier, called associate member, also was defined. Companies may apply to become associate members, who must pay membership fees. Adopter membership is free and open to anyone. In general, promoter and associate members develop and maintain the Bluetooth specification; adopter members may review specification updates before their public availability.

The SIG's original purpose was to develop the Bluetooth specification, but it has taken on additional responsibilities over time. In 2001, the SIG incorporated and instituted a more formal structure for the organization, including a board of directors that oversees all operations, a technical organization led by the Bluetooth Architecture Review Board (BARB), a marketing arm and a legal group. The SIG continues to develop and maintain the specification and promote the technology, including sponsoring developers conferences and other events. One important function of the SIG is to manage the Bluetooth qualification program, in which products are tested for conformance to the specification. All Bluetooth products must undergo qualification testing. The SIG's official Web site (Bluetooth SIG, 2002b) offers more details about the organization and the qualification program.

Wireless Personal Area Networks

A *personal area network* (PAN) generally is considered to be a set of communicating devices that someone carries with her. A wireless PAN (WPAN), of course, is such a set of devices that communicate without cables, such as through the use of Bluetooth technology. One can imagine a "sphere" of connectivity that surrounds a person and moves with her or him, so that all of the devices in the WPAN remain in communication with one another.

WPANs need only short-range communication capability to cover the personal area, in contrast with local area (LAN) or wide area networks (WAN), which need to communicate across greater distances using established infrastructure. One source (Miller, 2001) contrasts PANs, LANs, and WANs, particularly Bluetooth technology as a WPAN solution versus Institute of Electrical and Electronics Engineers (IEEE, 1999) 802.11 WLAN technology.

The usefulness of a WPAN derives primarily from the ability of individual devices to communicate with each other in an ad hoc manner. Each device still can specialize in certain capabilities but can "borrow" the capabilities of other devices to accomplish certain tasks. For example, a PDA is useful for quickly accessing personal information, such as appointments and contacts. A mobile telephone can be used to contact people whose information is stored in the PDA. Hence, a user might look up the telephone number of an associate and then dial that number on the mobile phone. With a WPAN, however, this process can be automated: Once the telephone number is accessed, the PDA software could include an option to dial the specified phone number automatically on the mobile telephone within the WPAN, using wireless communication links to transmit the dialing instructions to the phone. When combined with a wireless headset in the same WPAN, this could enable a more convenient device usage model for the user, who might never need to handle the mobile telephone at all (it could remain stored in a briefcase). Moreover, the user interface of the PDA is likely to be easier to use than a telephone keypad for retrieving contact information. This allows each of the devices (PDA, mobile phone, and wireless headset in this example) to be optimized to perform the specific tasks that they do best. The capabilities of each device are accessed from other devices via the WPAN. This often is preferred over an alternative usage model, the "all-in-one" device (imagine a PDA that also functions as a mobile telephone). Such multifunction devices might tend to be cumbersome and are more difficult to optimize for specific functions.

Because it was developed primarily to replace cables that connect mobile devices, Bluetooth wireless communication is an ideal WPAN technology. Indeed, most of the popular usage scenarios for Bluetooth technology originate in a WPAN of some sort, connecting personal devices to each other or to other networks in proximity. The use of Bluetooth communication links in WPANs is illustrated next, in an examination of various Bluetooth applications.

Bluetooth Applications

Because Bluetooth technology primarily is about replacing cables, many of its applications involve well-known usage scenarios. The value that Bluetooth communication adds to these types of applications derives from the ability to accomplish them without wires, enhancing mobility and convenience. For example, dial-up networking is a task commonly performed by many individuals, especially mobile professionals. One of the original usage scenarios used to illustrate the value of Bluetooth technology involves performing dial-up networking wirelessly—with the use of a mobile computer and a mobile phone, both equipped with Bluetooth communication, dial-up networking no longer is constrained by cables. This application and others are detailed next.

Basic Cable Replacement Applications

These applications comprise the original set of usage models envisioned for Bluetooth wireless technology. When the SIG was formed and the technology began to be popularized, these applications were touted as the most common ways in which Bluetooth wireless communication would be used. Although many other, perhaps more sophisticated, applications for Bluetooth technology have been discovered and envisioned, these original usage models remain as the primary application set. Nearly all of these early applications involve a mobile computer or a mobile telephone, and for the most part, they involve performing typical existing tasks wirelessly.

One such application, already mentioned, is dial-up networking. In this application, a Bluetooth communication link replaces the wire (typically a serial cable) between a computer and a telephone. When the telephone also is a mobile device, the network connection can be entirely wireless; a Bluetooth wireless link exists between the computer and the telephone, and a wide-area communications link (using typical cellular technology, such as the global system for mobile communication [GSM], time division/demand multiple access [TDMA], or others) carries the network traffic, using the mobile telephone as a wireless modem. Figure 1 depicts this usage model.

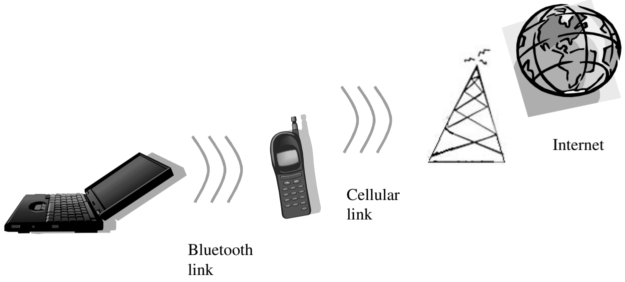


Figure 1: Dial-up networking illustration.

A variant of this application uses direct (rather than dial-up) connection to a network such as the Internet. In this case, the Bluetooth link allows a computer to connect to a network such as a LAN without using a cable. Together, these two applications (wireless dial-up networking and wireless LAN access) form a usage model that the SIG calls the *Internet Bridge*. Both applications involve access to a network, using existing protocols, with the main benefit being the ability to access the network without the cables that typically are required to connect the network client computer.

Another type of cable replacement application involves data transfer from one device to another. One of the most common such usage models is transferring files from one computer to another. This can be accomplished with removable media (diskettes, CDs), with cables (over a network or via a direct connection) or wirelessly (using infrared or Bluetooth communication, to name two ways). Infrared file transfer is not uncommon, but it requires the two devices to have a line of sight between them. Bluetooth file transfer operates similarly to that of IrDA® infrared file transfer (in fact, the Bluetooth protocol stack, discussed later, is designed such that the same application can be used over either transport medium). Bluetooth communication, being RF-based, does not require a line of sight between the two devices, however. Moreover, through the use of standard data formats, such as *vCard* (Internet Mail Consortium, 1996a), *vCal* (Internet Mail Consortium, 1996b), and others, objects other than files can be exchanged between devices using Bluetooth links in a manner similar to that used with IrDA. So, for example, electronic business cards, calendar appointments, and contact information can be shared wirelessly among devices.

Building on this capability to exchange data objects is the application that allows these same objects to be synchronized. This means that data sets on two devices reflect the same information at the point in time when they are synchronized. Hence, in addition to simply sending a copy of contact information or a calendar appointment from one device to another, the full address book or calendar can be synchronized between the two devices so that they have the same set of contacts or appointments. This allows a user to enter information on any convenient device and then have that information reflected on other devices by synchronizing with those devices. In addition to the benefit of performing these tasks wirelessly, by using standard protocols and data formats information can be exchanged easily among many kinds of devices. Specialized cables to connect two computers, or custom cradles to connect a PDA to a computer, are not needed once Bluetooth technology enters the picture. Instead, the same data can be exchanged and synchronized to and from notebook computers, PDAs, mobile phones, pagers, and other devices. This illustrates a hallmark of the value of Bluetooth technology: a single standard wireless link can replace many cables of various types, allowing devices that otherwise might not be able to be connected to communicate easily.

Another data transfer application is related to those just described, but it has a distinguished usage model because of the kind of data it transfers, namely, image data. The SIG calls this usage model the *instant postcard,* and it involves transferring pictures from one device to another. One reason that this application is separately described is because it involves the use of a digital camera. Today, when a camera captures new images, they typically are loaded onto a computer of some sort (or perhaps a television or similar video device) to be displayed. Through the use of Bluetooth wireless technology, this image transfer can be accomplished more easily, but once again, this standard form of wireless link enables the same data to be transferred to other types of devices. For example, rather than uploading photos to a computer, the photos might be transferred to a mobile phone. Even if the phone's display is not suitable for viewing the photo, it still could be e-mailed to someone who could view it on his computer or other e-mail device.

Until now this chapter has focused on applications involving data, but Bluetooth wireless technology also is designed to transport voice traffic (audio packets), and some of the cable replacement applications take advantage of this fact. The most notable of these is the wireless headset. Cabled headsets that connect to a mobile phone are widely used today to allow hands-free conversations. Bluetooth technology removes the cable from the headset to the telephone handset, enabling wireless operation that can allow the phone to be stowed away in a briefcase, pocket, or purse. In fact, as noted earlier, this particular application was the basis for the invention of Bluetooth technology. As with the previously discussed applications, however, once a standard wireless link is established, additional ways to connect other kinds of devices present themselves. For example, the same Bluetooth headset used with a mobile phone might also be used with a stationary telephone (again to allow hands-free operation and increased mobility), as well as with a computer (to carry audio traffic to and from the computer). Furthermore, although Bluetooth wireless communication was not originally designed to carry more complex audio traffic (such as digital music), advances are being made that will allow it to do so. With this capability, the same wireless headset also could be used with home entertainment systems, car audio systems, and personal music players. Hence, the Bluetooth SIG dubs this usage model the *ultimate headset.*

A variation on the wireless headset usage model is what the SIG calls the *speaking laptop.* In this application, Bluetooth links carry audio data in the same manner as for the headset application, but in this case the audio data is routed between a telephone and a notebook computer's speaker and microphone, rather than to a headset. One usage scenario enabled with this application is that of using the notebook computer as a speakerphone: A call made to or from a mobile telephone can be transformed into a conference call ("put on the speaker") by using the speaker and microphone built into nearly all portable (and desktop) computers.

Cordless telephony is another application for Bluetooth technology. With a Bluetooth voice access point, or cordless telephone base station, a standard cellular mobile telephone also can be used as a cordless phone in a home or office. The Bluetooth link carries the voice traffic from the handset to the base station, with the call then being carried over the normal wired telephone network. This allows mobile calls to be made without incurring cellular usage charges. In addition, two handsets can function as walkie-talkies, or an intercom system, using direct Bluetooth links between them, allowing two parties to carry on voice conversations in a home, office, or public space without any telephone network at all. Because a single mobile telephone can be used as a standard cellular phone, a cordless phone, and an intercom, the SIG calls this cordless telephony application the *three-in-one phone* usage model.

Additional Applications

Although Bluetooth wireless technology was developed especially for cable-replacement applications such as those just cited, many observers quickly realized that Bluetooth communication could be used in other ways, too. Here I describe a few of the many potential applications of Bluetooth technology, beginning with those that the SIG already is in the process of specifying.

The Bluetooth SIG focused primarily on the cable replacement applications already discussed in the version 1.x specifications that it released. The SIG also is developing additional profiles (detailed later) for other types of Bluetooth applications. Among these are more robust personal area proximity networking, human interface devices, printing, local positioning, multimedia, and automotive applications.

Bluetooth personal area networking takes advantage of the ability of Bluetooth devices to establish communication with each other based on their proximity to each other, so that ad hoc networks can be formed. The *Bluetooth Network Encapsulation Protocol* (BNEP) allows Ethernet packets to be transported over Bluetooth links, thus enabling many classic networking applications to operate in Bluetooth piconets (piconets are discussed at length in the section Bluetooth Operation). This capability extends the Bluetooth WPAN to encompass other devices. An example is the formation of an ad hoc Bluetooth network in a conference room with multiple meeting participants. Such a network could facilitate collaborative applications such as white-boarding, instant messaging, and group scheduling. Such applications could allow group editing of documents and scheduling of follow-up meetings, all in real time. Nonetheless, it should be noted that although this scenario resembles classic intranet- or Internet-style networking in some respects, Bluetooth personal area networking is not as robust a solution for true networking solutions as is a WLAN technology, such as IEEE 802.11 (described in the section Related Technologies).

Replacing desktop computer cables with Bluetooth communication links fundamentally is a cable replacement application (dubbed the *cordless computer* by the SIG), and this was one of the originally envisioned Bluetooth usage scenarios, but the original specifications did not fully address it. The new *Human Interface Device* (HID) specification describes how Bluetooth technology can be used in wireless computer peripherals such as keyboards, mice, joysticks, and so on. The Bluetooth printing profiles specify methods for wireless printing using Bluetooth communication, including "walk up and print" scenarios that allow immediate printing from any device, including mobile telephones and PDAs, as well as notebook computers, to any usable printer in the vicinity. This application of Bluetooth technology can obviate the need for specialized network print servers and their associated configuration and administration tasks.

Another application that can be realized with Bluetooth wireless technology is *local positioning.* Bluetooth technology can be used to augment other technologies, such as global positioning systems (GPS), especially inside buildings, where other technologies might not work well. Using two or more Bluetooth radios, local position information can be obtained in several ways. If one Bluetooth device is stationary (say, a kiosk), it could supply its position information to other devices within range. Any device that knows its own position can provide this information to other Bluetooth devices so that they can learn their current position. Sophisticated applications might even use signal strength information to derive more granular position information. Once position information is known, it could be used with other applications, such as maps of the area, directions to target locations, or perhaps even locating lost devices. *The Bluetooth Local Positioning Profile* specifies standard data formats and interchange methods for local positioning information.

Multimedia applications have become standard on most desktop and notebook computers, and the Bluetooth SIG is pursuing ways by which streamed multimedia data, such as sound and motion video, could be used in Bluetooth environments. The 1 Mbps raw data rate of version 1 Bluetooth radios is not sufficient for many sorts of multimedia traffic, but the SIG is investigating methods for faster Bluetooth radios that could handle multimedia applications. The SIG also is developing profiles for multimedia data over Bluetooth links.

Another emerging application area is that of automotive Bluetooth applications. Using Bluetooth communication, wireless networks could be formed in cars. Devices from the WPAN could join the automobile's built-in Bluetooth network to accomplish scenarios such as the following:

- Obtaining e-mail and other messages, using a mobile phone as a WAN access device, and transferring those messages to the car's Bluetooth network, where they might be read over the car's audio system using text-to-speech technology (and perhaps even composing responses using voice recognition technology)
- Obtaining vehicle information remotely, perhaps for informational purposes (for example, querying the car's current mileage from the office or home) or for diagnostic purposes (for example, a wireless engine diagnostic system for automobile mechanics that does not require probes and cables to be connected to the engine)
- Sending alerts and reminders from the car to a WPAN when service or maintenance is required (for example, e-mail reminders that an oil change is due or in-vehicle or remote alerts when tire pressure is low or other problems are diagnosed by the vehicle's diagnostic systems)

The SIG, in conjunction with automotive industry representatives, is developing profiles for Bluetooth automotive applications. This area is likely to prove to be an exciting and rapidly growing domain for the use of Bluetooth wireless technology.

These new applications comprise only some of the potential uses for Bluetooth wireless technology. Many other domains are being explored or will be invented in the future. Other noteworthy applications for Bluetooth wireless communications include mobile e-commerce, medical, and travel technologies. Bluetooth devices such as mobile phones or PDAs might be used to purchase items in stores or from vending machines; wireless biometrics and even Bluetooth drug dispensers might appear in the future (the 2.4-GHz band in which Bluetooth operates is called the industrial, scientific, and medical band); and travelers could experience enhanced convenience by using Bluetooth devices for anytime, anywhere personal data access and airline and hotel automated check-in. In fact, this latter scenario, including the use of a Bluetooth PDA to unlock a hotel room door, has already been demonstrated (InnTechnology, 2000).

The Bluetooth Protocol Stack

A complete discussion of the Bluetooth protocol stack is outside the scope of this article. Numerous books, including Miller and Bisdikian (2000) and Bray and Sturman (2000) offer more in-depth discussions of Bluetooth protocols. Here we present an overview of Bluetooth operation and how the various protocols may be used to accomplish the applications already discussed. A typical Bluetooth stack is illustrated in Figure 2. Each layer of the stack is detailed next.

Radio, Baseband, and Link Manager

These three protocol layers comprise the *Bluetooth module.* Typically, this module is an electronics package containing hardware and firmware. Today many manufacturers supply Bluetooth modules.

The *radio* consists of the signal processing electronics for a transmitter and receiver (transceiver) to allow RF communication over an *air-interface* between two Bluetooth devices. As noted earlier, the radio operates in the 2.4-GHz spectrum, specifically in the frequency range 2.400–2.4835 GHz. This frequency range is divided into

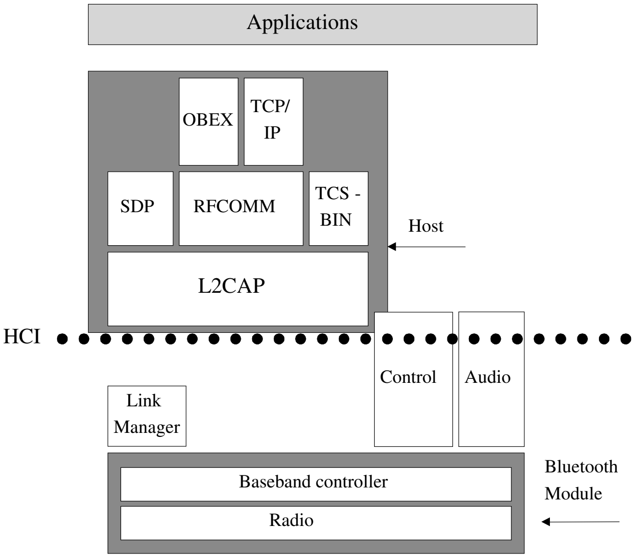


Figure 2: The Bluetooth protocol stack.

79 channels (along with upper and lower guard bands), with each channel having a 1-MHz separation from its neighbors. *Frequency hopping* is employed in Bluetooth wireless communication; each packet is transmitted on a different channel, with the channels being selected pseudo-randomly, based on the clock of the master device (master and slave devices are described in more detail later). The receiving device knows the frequency hopping pattern and follows the pattern of the transmitting device, hopping to the next channel in the pattern to receive the transmitted packets.

The Bluetooth specification defines three classes of radios, based on their maximum power output:

- 1 mW (0 dBm)
- 2.5 mW (4 dBm)
- 100 mW (20 dBm)

Increased transmission power offers a corresponding increase in radio range; the nominal range for the 0-dBm radio is 10 m, whereas the nominal range for the 20-dBm radio is 100 m. Of course, increased transmission power also requires a corresponding increase in the energy necessary to power the system, so higher power radios will draw more battery power. The basic cable replacement applications (indeed, most Bluetooth usage scenarios described here) envision the 0-dBm radio, which is considered the standard Bluetooth radio and is the most prevalent in devices. The 0-dBm radio is sufficient for most applications, and its low power consumption makes it suitable for use on small, portable devices.

Transmitter and receiver characteristics such as interference, tolerance, sensitivity, modulation, and spurious emissions are outside the scope of this chapter but are detailed in the Bluetooth specification (Bluetooth SIG, 2001a).

The *baseband controller* controls the radio and typically is implemented as firmware in the Bluetooth module. The controller is responsible for all of the various timing and raw data handling aspects associated with RF communication, including the frequency hopping just mentioned, management of the time slots used for transmitting and receiving packets, generating air-interface packets (and causing the radio to transmit them), and parsing air-interface packets (when they are received by the radio). Packet generation and reception involves many considerations, including the following:

- Generating and receiving packet payload
- Generating and receiving packet headers and trailers
- Dealing with the several packet formats defined for Bluetooth communication
- Error detection and correction
- Address generation and detection
- Data whitening (a process by which the actual data bits are rearranged so that the occurrence of zero and one bits in a data stream is randomized, helping to overcome DC bias)
- Data encryption and decryption

Not all of these operations are necessarily performed on every packet; there are various options available for whether or not a particular transformation is applied to the data, and in some cases (such as error detection and correction), there are several alternatives that may be employed by the baseband firmware.

The *link manager,* as its name implies, manages the link layer between two Bluetooth devices. Link managers in two devices communicate using the *link manager protocol* (LMP). LMP consists of a set of commands and responses to set up and manage a baseband link between two devices. A link manager on one device communicates with a link manager on another device (indeed, this is generally the case for all the Bluetooth protocols described here; a particular protocol layer communicates with its corresponding layer in the other device, using its own defined protocol. Each protocol is passed to the next successively lower layer, where it is transformed to that layer's protocol, until it reaches the baseband, where the baseband packets that encapsulate the higher layer packets are transmitted and received over the air interface). LMP setup commands include those for authenticating the link with the other device; setting up encryption, if desired, between the two devices; retrieving information about the device at the other end of the link, such as its name and timing parameters; and swapping the master and slave roles (detailed later) of the two devices. LMP management commands include those for controlling the transmission power; setting special power-saving modes, called hold, park, and sniff; and managing quality-of-service (QoS) parameters for the link. Because LMP messages deal with fundamental characteristics of the communication link between devices, they are handled in an expedited manner, at a higher priority than the normal data that is transmitted and received.

Control and Audio

The *control* and *audio* blocks in Figure 2 are not actual protocols. Instead, they represent means by which the upper layers of the stack can access lower layers. The control functions can be characterized as methods for inter-protocol-layer communication. These could include requests and notifications from applications, end users, or protocol layers that require action by another protocol layer, such as setting desired QoS parameters, requests to enter or terminate power-saving modes, or requests to search for other Bluetooth devices or change the discoverability of the local device. Often, these take the form of a user-initiated action, via an application, that requires the link manager (and perhaps other layers) to take some action.

The audio block in Figure 2 represents the typical path for audio (voice) traffic. Recall that Bluetooth wireless technology supports both data and voice. Data packets traverse through the L2CAP layer (described later), but voice packets typically are routed directly to the baseband, because audio traffic is isochronous and hence time critical. Audio traffic usually is associated with telephony applications, for which data traffic is used to set up and control the call and voice traffic serves as the content of the call. Audio data can be carried over Bluetooth links in two formats:

- Pulse code modulation (PCM), with either a-law or μ-law logarithmic compression
- Continuous variable slope delta (CVSD) modulation, which works well for audio data with relatively smooth continuity, usually the case for typical voice conversations

Host-Controller Interface (HCI)

The HCI is an optional interface between the two major components of the Bluetooth stack: the *host* and the *controller*. As shown in Figure 2 and described earlier, the radio, baseband controller, and link manager comprise the module, which is often, but not necessarily, implemented in a single electronics package. Such a module can be integrated easily into many devices, with the remaining layers of the stack residing on the main processor of the device (such as a notebook computer, mobile phone, or PDA). These remaining layers (described next) are referred to as the host portion of the stack.

Figure 2 illustrates a typical "two-chip" solution in which the first "chip" is the Bluetooth module and the second "chip" is the processor in the device on which the host software executes. (The module itself might have multiple chips or electronic subsystems for the radio, the firmware processor, and other external logic.) In such a system, the HCI allows different Bluetooth modules to be interchanged in a device, because it defines a standard method for the host software to communicate with the controller firmware that resides on the module. So, at least in theory, one vendor's Bluetooth module could be substituted for another, so long as both faithfully implement the HCI. Although this is not the only type of partitioning that can be used when implementing a Bluetooth system, the SIG felt it was common enough that a standard interface should be defined between the two major components of the system. A Bluetooth system could be implemented in an "all-in-one" single module, where the host and controller reside together in the same physical package (often called a "single-chip" solution), although in this case, the HCI might still be used as an internal interface. When the two-chip solution is used, the physical layer for the HCI (that is, the physical connection between the host and the controller) could be one of several types. The Bluetooth specification defines three particular physical layers for the HCI:

- Universal serial bus (USB)
- Universal asynchronous receiver/transmitter (UART)
- RS-232 serial port

Other HCI transports could be implemented; the Bluetooth specification currently contains details and considerations for these three.

Logical Link Control and Adaptation Protocol (L2CAP)

The Logical Link Control and Adaptation Protocol (L2CAP) layer serves as a "funnel" through which all data traffic flows. As discussed earlier, voice packets typically are routed directly to the baseband, whereas data packets flow to and from higher layers, such as applications, to the baseband via the L2CAP layer.

The L2CAP layer offers an abstraction of lower layers to higher layer protocols. This allows the higher layers to operate using more natural data packet formats and protocols, without being concerned about how their data is transferred over the air-interface. For example, the Service Discovery Protocol layer (discussed next) defines its own data formats and protocol data units. At the SDP layer, only the service discovery protocol needs to be handled; the fact that SDP data must be separated into baseband packets for transmission and aggregated from baseband packets for reception is not a concern at the SDP layer, nor are any of the other operations that occur at the baseband (such as encryption, whitening, and so on). This is accomplished because the L2CAP layer performs operations on data packets. Among these operations are segmentation and reassembly, whereby the L2CAP layer breaks higher layer protocol data units into L2CAP packets, which in turn can be transformed into baseband packets; the L2CAP layer conversely can reassemble baseband packets into L2CAP packets that in turn can be transformed into the natural format of higher layers of the stack.

An L2CAP layer in one Bluetooth stack communicates with another, corresponding L2CAP layer in another Bluetooth stack. Each L2CAP layer can have many channels. L2CAP channels identify data streams between the L2CAP layers in two Bluetooth devices. (L2CAP channels should not be confused with baseband channels used for frequency hopping. L2CAP channels are logical identifiers between two L2CAP layers.) An L2CAP channel often is associated with a particular upper layer of the stack, handling data traffic for that layer, although there need not be a one-to-one correspondence between channels and upper-layer protocols. An L2CAP layer might use the same protocol on multiple L2CAP channels. This illustrates another data operation of the L2CAP layer: protocol multiplexing. Through the use of multiple channels and a protocol identifier (called a *protocol-specific* *multiplexer,* or PSM), L2CAP allows various protocols to be multiplexed (flow simultaneously) over the air-interface. The L2CAP layer sorts out which packets are destined for which upper layers of the stack.

Service Discovery Protocol (SDP)

The SDP layer provides a means by which Bluetooth devices can learn, in an ad hoc manner, which services are offered by each device. Once a connection has been established, devices use the SDP to exchange information about services. An SDP client queries an SDP server to inquire about services that are available; the SDP server responds with information about services that it offers. Any Bluetooth device can be either an SDP client or an SDP server, acting in one role or the other at different times.

SDP allows a device to inquire about specific services in which it is interested (called *service searching*) or to perform a general inquiry about any services that happen to be available (called *service browsing*). A device can perform an SDP service search to look for, say, printing services in the vicinity. Any devices that offer a printing service that matches the query can respond with a "handle" for the service; the client then uses that handle to perform additional queries to obtain more details about the service. Once a service is discovered using SDP, other protocols are used to access and invoke the service; one of the items that can be discovered using SDP is the set of protocols that are necessary to access and invoke the service.

SDP is designed to be a lightweight discovery protocol that is optimized for the dynamic nature of Bluetooth piconets. SDP can coexist with other discovery and control protocols; for example, the Bluetooth SIG has published a specification for using the UPnP discovery and control technology over Bluetooth links.

RFCOMM

As its name suggests, the RFCOMM layer defines a standard communications protocol, specifically one that emulates serial port communication (the "RF" designates radio frequency wireless communication; the "COMM" portion suggests a serial port, commonly called a COM port in the personal computer realm). RFCOMM emulates a serial cable connection and provides the abstraction of a serial port to higher layers in the stack. This is particularly valuable for Bluetooth cable replacement applications, because so many cable connections—modems, infrared ports, camera and mobile phone ports, printers, and others—use some form of a serial port to communicate.

RFCOMM is based on the European Telecommunications Standards Institute (ETSI) TS07.10 protocol (ETSI, 1999), which defines a multiplexed serial communications channel. The Bluetooth specification adopts much of the TS07.10 protocol and adds some Bluetooth adaptation features. The presence of RFCOMM in the Bluetooth protocol stack is intended to facilitate the migration of existing wired serial communication applications to wireless Bluetooth links. By presenting higher layers of the stack with a virtual serial port, many existing applications that already use a serial port can be used in Bluetooth environments without any changes. Indeed, many of the cable replacement applications cited earlier, including dial-up networking, LAN access, headset, and file and object exchange, use RFCOMM to communicate. Because RFCOMM is a multiplexed serial channel, many serial data streams can flow over it simultaneously; each separate serial data stream is identified with a *server channel,* in a manner somewhat analogous to the channels used with L2CAP.

Telephony Control Specification-Binary (TCS-BIN)

The TCS-BIN is a protocol used for advanced telephony operations. Many of the Bluetooth usage scenarios involve a mobile telephone, and some of these use the TCS-BIN protocol. TCS-BIN is adopted from the ITU-T Q.931 standard (International Telecommunication Union, 1998), and it includes functions for call control and managing wireless user groups. Typically, TCS-BIN is used to set up and manage voice calls; the voice traffic that is the content of the call is carried over audio packets as described earlier. Applications such as the three-in-one phone usage model use TCS-BIN to enable functions such as using a mobile phone as a cordless phone or an intercom. In these cases, TCS-BIN is used to recognize the mobile phone so that it can be added to a wireless user group that consists of all the cordless telephone handsets used with a cordless telephone base station. TCS-BIN also is used to set up and control calls between the handset and the base station (cordless telephony) or between two handsets (intercom).

TCS-BIN offers several advanced telephony functions; devices that support TCS-BIN can obtain knowledge of and directly communicate with any other devices in the TCS-BIN wireless user group, essentially overcoming the master-slave relationship of the underlying Bluetooth piconet (detailed later). It should be noted that not all Bluetooth telephony applications require TCS-BIN; an alternative method for call control is the use of AT commands over the RFCOMM serial interface. This latter method is used for the headset, dial-up networking, and fax profiles.

Adopted Protocols

Although several layers of the Bluetooth protocol stack were developed specifically to support Bluetooth wireless communication, other layers are adopted from existing industry standards. I already have noted that RFCOMM and TCS-BIN are based on existing specifications. In addition to these, protocols for file and object exchange and synchronization are adopted from the Infrared Data Association (IrDA), and Internet networking protocols are used in some applications.

The IrDA's *object exchange* (OBEX) protocol is used for the file and object transfer, object push, and synchronization usage models. OBEX originally was developed for infrared wireless communication, and it maps well to Bluetooth wireless communication. OBEX is a relatively lightweight protocol for data exchange, and several well-defined data types—including electronic business cards, e-mail, short messages, and calendar items—can be carried within the protocol. Hence, the Bluetooth SIG adopted OBEX for use in its data exchange scenarios; by doing so, existing infrared applications can be used over Bluetooth links, often with no application changes.

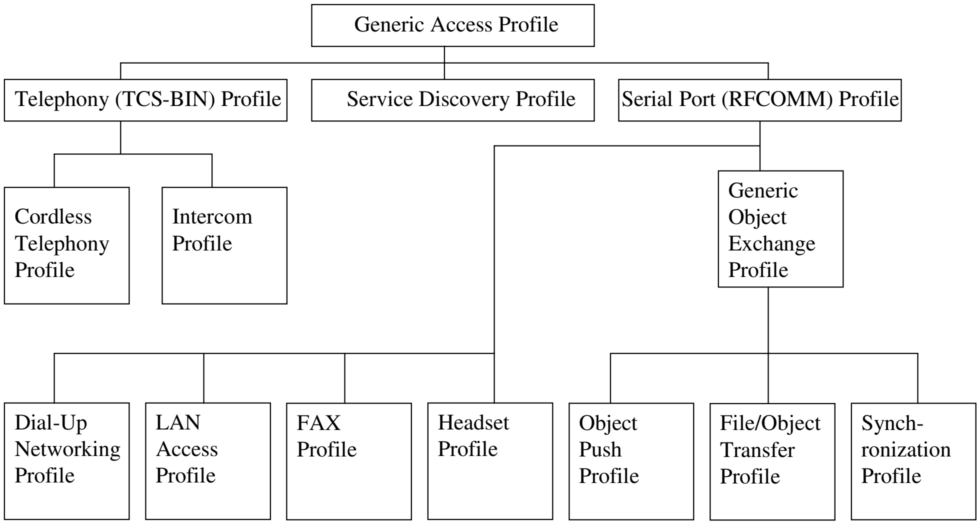

Generic Access Profile
Telephony (TCS-BIN) Profile
Service Discovery Profile
Serial Port (RFCOMM) Profile
Cordless Telephony Profile
Intercom Profile
Generic Object Exchange Profile
Dial-Up Networking Profile
LAN Access Profile
FAX Profile
Headset Profile
Object Push Profile
File/Object Transfer Profile
Synch-ronization Profile

Figure 3: The Bluetooth profiles.

In addition, the *infrared mobile communications* (IrMC) protocol is used for the synchronization usage model. Typically, infrared communication occurs over a serial port, so the adopted IrDA protocols operate over RFCOMM in the Bluetooth protocol stack.

Networking applications such as dial-up networking and LAN access use standard Internet protocols, including point-to-point protocol (PPP), Internet protocol (IP), user datagram protocol (UDP), and transmission control protocol (TCP). As shown in Figure 2, these protocols operate over the RFCOMM protocol. Once a Bluetooth RFCOMM connection is established between two devices, PPP can be used as a basis for UDP–IP and TCP–IP networking packets. This enables typical networking applications, such as network dialers, e-mail programs, and browsers, to operate over Bluetooth links, often with no changes to the applications.

Bluetooth Profiles

I have presented an overview of the Bluetooth protocols, which are specified in Bluetooth SIG (2001a), the first volume of the Bluetooth specification. The Bluetooth SIG also publishes a second volume of the specification (Bluetooth SIG, 2001b), which defines the Bluetooth *profiles*. Profiles offer additional guidance to developers beyond the specification of the protocols. Essentially, a profile is a formalized usage case that describes how to use the protocols (including which protocols to use, which options are available, and so on) for a given application. Profiles were developed to foster interoperability; they provide a standard basis for all implementations, to increase the likelihood that implementations from different vendors will work together, so that end users can have confidence that Bluetooth devices will interoperate with each other. In addition to the profile specifications, the SIG offers other mechanisms intended to promote interoperability; among these are the *Bluetooth Qualification Program* (a definition of testing that a Bluetooth device must undergo) and *unplugfests* (informal sessions where many vendors can test their products with each other); detailed discussions of these programs are outside the scope of this chapter, but more information is available on the Bluetooth Web site (Bluetooth SIG, 2002b).

Our earlier discussion of Bluetooth applications presented several usage models for Bluetooth wireless communication. Many of these applications have associated profiles. For example, the dial-up networking profile defines implementation considerations for the dial-up networking application. Most of the applications cited here have associated profiles, and many new profiles are being developed and published by the SIG; the official Bluetooth Web site (Bluetooth SIG, 2002b) has a current list of available specifications. In addition, there are some fundamental profiles that describe basic Bluetooth operations that are necessary for most any application. The version 1.1 profiles are illustrated in Figure 3. This figure shows the relationship among the various profiles, illustrating how certain profiles are derived from (and build upon) others.

The leaf nodes of the diagram consist of profiles that describe particular applications—file and object transfer, object push, synchronization, dial-up networking, fax, headset, LAN access, cordless telephony, and intercom. The telephony (TCS-BIN) profile includes elements that are common to cordless telephony and intercom applications; similarly, the generic object exchange profile describes the common elements for its children, and the serial port profile defines operations used by all applications that use the RFCOMM serial cable replacement protocol. Note that the generic object exchange profile derives from the serial port profile; this is because OBEX operates over RFCOMM in the Bluetooth protocol stack.

The two remaining profiles describe fundamental operations for Bluetooth communication. The service discovery application profile describes how a service discovery application uses the service discovery protocol (described earlier). The generic access profile is common to all applications; it defines the basic operations that Bluetooth devices use to establish connections, including how devices become discoverable and connectable, security considerations for connections, and so on. The generic access profile also includes a common set of terminology used in other profiles; this is intended to reduce ambiguity in the specification. The generic access and service discovery application profiles are mandatory for all Bluetooth devices to implement, because they form the basis for interoperable devices. Other works, including Miller and Bisdikian (2000), delve more deeply into the Bluetooth profiles.

Bluetooth Operation

Having discussed WPANs, Bluetooth applications, protocols, and profiles, I now turn our attention to some of the fundamental concepts of Bluetooth operation, illustrating an example flow for a Bluetooth connection.

At the baseband layer, Bluetooth operates on a master–slave model. In general, the *master* device is the one that initiates communication with one or more other devices. *Slaves* are the devices that respond to the master's queries. In general, any Bluetooth device can operate as either a master or a slave at any given time. The master and slave roles are meaningful only at the baseband layer; upper layers are not concerned with these roles. The master device establishes the frequency hopping pattern for communication with its slaves, using its internal clock values to generate the frequency hopping pattern. Slaves follow the frequency hopping pattern of the master(s) with which they communicate.

When a master establishes a connection with one or more slaves, a *piconet* is formed. To establish the connection, a master uses processes called *inquiry* and *paging.* A master can perform an inquiry operation, which transmits a well-defined data sequence across the full spectrum of frequency hopping channels. An inquiry effectively asks, "Are there any devices listening?" Devices that are in *inquiry scan* mode (a mode in which the device periodically listens to all of the channels for inquiries) can respond to the master's inquiry with enough information for the master device to address the responding device directly. The inquiring (master) device may then choose to page the responding (slave) device. The page also is transmitted across the full spectrum of frequency hopping channels; the device that originally responded to the inquiry can enter a *page scan* state (a state in which it periodically listens to all of the channels for pages), and it can respond to the page with additional information that can be used to establish a baseband connection between the master and the slave. The master can repeat this process and establish connections with as many as seven slaves at a time. Hence, a piconet consists of one master and up to 7 active slaves; additional slaves can be part of the piconet, but only seven slaves can be active at one time. Slaves can be "parked" (made inactive) so that other slaves can be activated. Figure 4 illustrates a typical Bluetooth piconet. Note that a device could be a slave in more than one piconet at a time, or it could be a master of one piconet and a slave in a second piconet. In these cases, the device participating in multiple piconets must use the appropriate frequency hopping pattern in each piconet, so it effectively must split its time among all the piconets in which it participates. The Bluetooth specification calls such interconnected piconets *scatternets*.

Once a piconet is formed (a baseband connection exists between a master and one or more slave devices), higher layer connections can be formed, and link manager commands and responses may be used to manage the link. At some point, it is likely that an L2CAP connection will be formed for data packets (even if the main content for a link is voice traffic, an L2CAP data connection will be needed to set up and manage the voice links). A Bluetooth device can have one data (L2CAP) connection and up to three voice connections with any other Bluetooth device at a given time (recall, however, that L2CAP connections can be multiplexed, so many different data streams can flow over the single L2CAP connection). Once an L2CAP connection is established, additional higher layer connections can be established. If the devices are not familiar to each other already, it is likely that an SDP connection will be used to perform service discovery. RFCOMM and TCS-BIN connections also might be made, depending on the application. From here, applications can manage voice and data packets to accomplish their usage scenarios, which might be those defined by Bluetooth profiles or other ways to use Bluetooth wireless communication to accomplish a given task. Additional details about fundamental Bluetooth operations and connection establishment, including the various types of packets, master–slave communication protocols, and timing considerations, are outside the scope of this chapter but are detailed in works such as Miller and Bisdikian (2000) and Bray and Sturman (2000).

An additional noteworthy aspect of Bluetooth operation is security. In the wireless world, security justifiably is a key concern for device manufacturers, device deployers and end users. The Bluetooth specification includes security measures such as authentication and encryption. At the time that a link is established, Bluetooth devices may be required to authenticate themselves to each other. Once a link has been established, the data traffic over that

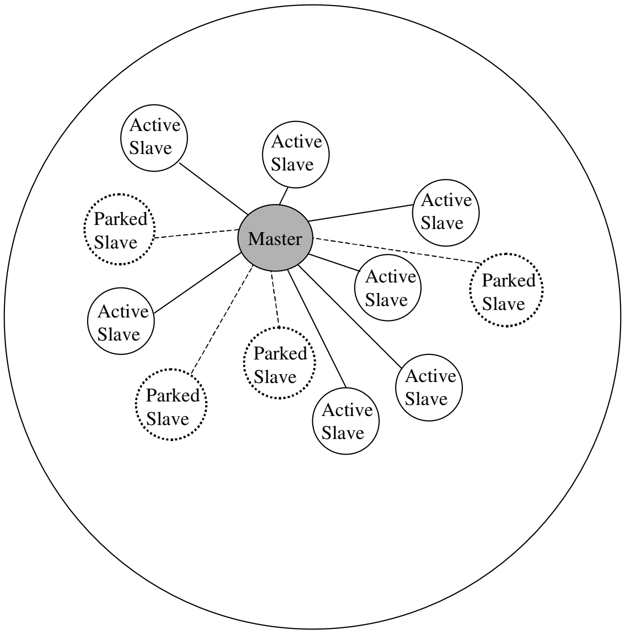


Figure 4: Example of a Bluetooth piconet.

link may be required to be encrypted. The methods used for authentication and encryption are detailed in the specification, and the Bluetooth profiles discuss what security measures should be employed in various circumstances (for example, the file transfer profile requires that both authentication and encryption be supported, and it recommends that they be used). Applications are free to impose additional security restrictions beyond those that are provided in the specification. For example, applications might choose to expose data or services only to authorized users or to implement more robust user authentication schemes. Details of the operation of Bluetooth security features are outside the scope of this chapter but are detailed in works such as Miller and Bisdikian (2000) and Bray and Sturman (2000).

Related Technologies

Other sources, including other chapters in this encyclopedia (see the "Cross References" section), discuss related wireless communication technologies in some depth. Here I briefly describe two particularly interesting related technologies, IrDA and IEEE 802.11 WLANI. I also comment on the IEEE 802.15.1 WPAN standard.

IrDA technology (IrDA 2002) uses infrared optical links, rather than RF links, to accomplish wireless communication between two devices. IrDA ports are common on many portable computing and communication devices and often are found on the same sorts of devices that are good candidates for using Bluetooth technology. IrDA and Bluetooth technologies share several aspects:

- Both are for short-range communication.
- Both use low-power transmission.
- Both are useful for cable-replacement applications.

Some differences between the two technologies are the following:

- IrDA typically uses less power than Bluetooth radios.
- IrDA data rates typically are greater than those of Bluetooth wireless communication.
- IrDA technology, because it uses optical communication, requires a "line of sight" between the two devices, whereas Bluetooth technology, because it uses RF communication, can penetrate many obstacles that might lie between two devices.

As already noted, many IrDA applications are well suited for use with Bluetooth links, and the Bluetooth SIG adopted the OBEX and IrMC protocols from the IrDA specification for use in the Bluetooth specification to foster the use of common applications for either technology. For certain applications in certain situations, one technology or the other might be the most suitable to use. A more detailed comparison of the two technologies can be found in Suvak (1999).

Many articles (e.g., Miller, 2001) have dealt with the relationship between IEEE 802.11 WLAN (IEEE, 1999) and Bluetooth WPAN technologies. Often, these two technologies are portrayed as being in competition with one another, with the proposition that one will "win" at the expense of the other. In fact, I believe, as many do, that these two technologies are complementary. Indeed, they are optimized for different purposes and applications. IEEE 802.11 is for WLAN applications, whereas Bluetooth technology is aimed at WPAN applications. Similarities between the two technologies include the following:

- Both operate in the 2.4-GHz RF spectrum. (The IEEE 802.11 standard actually includes two variants, 802.11a and 802.11b. Here I focus primarily on 802.11b. IEEE 802.11a operates in the 5-GHz spectrum.)
- Both use spread spectrum to communicate.
- Both can be used to access networks.

There are, however, distinct differences between the two technologies, including the following:

- IEEE 802.11 has a longer range (nominally 100 m) versus the typical nominal range of 10 m for Bluetooth wireless communication.
- IEEE 802.11 permits a much faster data rate (about 11 Mbps for 802.11b, even faster for 802.11a) than does Bluetooth technology (about 1 Mbps).
- The topology, higher data rate, and longer range of IEEE 802.11 WLAN technology inevitably lead to higher power consumption on average, making it less suitable for small portable personal devices with limited battery power.
- IEEE 802.11 is intended to be used for Ethernet-style networking applications in a LAN environment—it is a typical LAN without the wires—whereas Bluetooth is optimized for WPAN applications, notably cable replacement.

Indeed, the IEEE distinguishes between WLAN and WPAN technologies and has a separate standard, IEEE 802.15.1 (IEEE, 2001) for WPAN applications. In fact, this IEEE 802.15.1 standard is based on the Bluetooth specification, and it essentially adopts a subset of the Bluetooth technology as the IEEE standard. Using Bluetooth wireless technology as the basis for the IEEE WPAN standards offers further evidence that Bluetooth and IEEE 802.11 technologies can complement each other, as does the fact that the Bluetooth SIG and the IEEE work together on certain issues, including pursuing methods by which the RF interference between the two technologies can be minimized.

CONCLUSION

This introduction to Bluetooth wireless technology has touched on what can be done with the technology (Bluetooth applications), how it began and how it now is managed (the Bluetooth SIG), how it works (Bluetooth protocols, profiles, and operation), and how it relates to some other wireless communication technologies. The chapter focused on the application of Bluetooth wireless communication as a WPAN technology for connecting personal portable devices. I have presented several references where this topic is explored in greater detail.

With tremendous industry backing, a design to work with both voice and data, the ability to replace cumbersome cables, and many new products being deployed on a regular basis, Bluetooth wireless technology is poised to become an important way for people to communicate for the foreseeable future. From its genesis as a method to provide a wireless headset for mobile phones, this technology named for a Danish king continues to spread across the planet.

GLOSSARY

Bluetooth wireless technology Name given to a wireless communications technology used for short-range voice and data communication, especially for cable-replacement applications.

Bluetooth SIG The Bluetooth Special Interest Group, an industry consortium that develops, promotes, and manages Bluetooth wireless technology, including the Bluetooth qualification program and the Bluetooth brand.

Frequency-hopping spread spectrum A method of dividing packetized information across multiple channels of a frequency spectrum that is used in Bluetooth wireless communication.

IEEE 802.11 A wireless local-area network standard developed by the Institute for Electrical and Electronics Engineers that is considered complementary to Bluetooth wireless technology.

IEEE 802.15.1 A wireless personal area network standard developed by the Institute for Electrical and Electronics Engineers that is based on Bluetooth wireless technology.

Infrared Data Association (IrDA) An industry consortium that specifies the IrDA infrared communication protocols, some of which are used in the Bluetooth protocol stack.

Piconet A Bluetooth wireless technology term for a set of interconnected devices with one master and up to seven active slave devices.

Profile In Bluetooth wireless technology, a specification for standard methods to use when implementing a particular application, with a goal of fostering interoperability among applications and devices.

Radio frequency (RF) Used in the Bluetooth specification to describe the use of radio waves for physical layer communication.

Wireless personal area network (WPAN) A small set of interconnected devices used by one person.

CROSS REFERENCES

See *Mobile Commerce; Mobile Devices and Protocols; Mobile Operating Systems and Applications; Propagation Characteristics of Wireless Channels; Radio Frequency and Wireless Communications; Wireless Application Protocol (WAP); Wireless Communications Applications; Wireless Internet.*

REFERENCES

Bluetooth Special Interest Group (2001a). *Specification of the Bluetooth system* (Vol. 1). Retrieved December 10, 2002, from http://www.bluetooth.com/pdf/Bluetooth_11_Specifications_Book.pdf

Bluetooth Special Interest Group (2001b). *Specification of the Bluetooth system*, (Vol. 2). Retrieved December 10, 2002, from http://www.bluetooth.com/pdf/Bluetooth_11_Specifications_Book.pdf

Bluetooth Special Interest Group (2002a). *Trademark info*. Retrieved December 10, 2002, from http://www.bluetooth.com/sig/trademark.use.asp

Bluetooth Special Interest Group (2002b). *The official Bluetooth Web site*. Retrieved December 10, 2002, from http://www.bluetooth.com

Bray, J. & Sturman, C. (2000) *Bluetooth: Connect without cables*. New York: Prentice Hall PTR. (Second edition published 2001.)

European Telecommunications Standards Institute (1999). *Technical specification: Digital cellular telecommunications system (Phase 2+); Terminal equipment to mobile station (TE-MS) multiplexer protocol* (GSM 07.10). Retrieved March 28, 2003, from http://www.etsi.org

Infrared Data Association (2002). *IrDA SIR data specification* (and related documents). Retrieved March 28, 2003, from http://www.irda.org/standards/specifications.asp

InnTechnology (2000). *The Venetian Resort-Hotel-Casino & InnTechnology showcase Bluetooth hospitality services*. Retrieved March 28, 2003, from http://www.inntechnology.com/bluetooth/bluetooth_press_release.html

Institute of Electrical and Electronics Engineers (1999). *Wireless Standards Package (802.11)*. Retrieved March 28, 2003, from http://standards.ieee.org/getieee802

Institute of Electrical and Electronics Engineers (2001). *IEEE 802.15 Working Group for WPANs*. Retrieved March 28, 2003, from http://standards.ieee.org/getieee802

International Telecommunication Union (1998). *Recommendation Q.931—ISDN user-network interface layer 3 specification for basic call control*. Retrieved March 28, 2003, from http://www.itu.org

Internet Mail Consortium (1996a). *vCard—The electronic business card exchange format*. Retrieved March 28, 2003, from http://www.imc.org/pdi

Internet Mail Consortium (1996b). *vCalendar—The electronic calendaring and scheduling exchange format*. Retrieved March 28, 2003, from http://www.imc.org/pdi

Kardach, J. (2001). The naming of a technology. *Incisor, 34* (10–12) and *37* (13–15).

Miller, B. (2001). *The phony conflict: IEEE 802.11 and Bluetooth wireless technology.* IBM DeveloperWorks. Retrieved March 28, 2003, from http://www-106.ibm.com/developerworks/library/wi-phone/?dwzone = wireless

Miller, B., & Bisdikian, C. (2000). *Bluetooth revealed: The insider's guide to an open specification for global wireless communication*. New York: Prentice Hall PTR. (Second edition published 2001.)

Suvak, D. (1999). *IrDA and Bluetooth: A complementary comparison*. Walnut Creek, CA: Infrared Data Association. Retrieved March 28, 2003, from http://www.irda.org/design/ESIIrDA_Bluetoothpaper.doc

Business Plans for E-commerce Projects

Amy W. Ray, *Bentley College*

INTRODUCTION AND BACKGROUND

Numerous excellent resources exist to help would-be entrepreneurs with the technical aspects of writing a business plan, many of which are referenced here. Many articles have also been written on financial management of e-businesses, but most of these are targeted to helping venture capitalists (VCs) rather than entrepreneurs. Would-be entrepreneurs should begin by taking one step back from the actual preparation of a written business plan to consider the types of funding available for start-up e-businesses. Deciding how best to fund a start-up company is the first important issue faced by entrepreneurs, yet the consequences of specific choices are often overlooked—the people who fund the company will invariably have a major impact on how the company is ultimately managed.

Although there are numerous benefits to actually writing a business plan, most formal business plans are executed as a requirement for obtaining external funding, and the primary source of funding for e-businesses is venture capital. In fact, Venture Economics, a New York-based consulting firm, notes that about $240 billion has been poured into venture funds since the beginning of 1998 (Healy, 2002). This number does not include the significant contributions from VCs beyond the large, centrally managed venture funds. Accordingly, the benefits and challenges of using venture capital as a primary source of start-up funding are discussed here. Through consulting engagements with numerous start-up companies, the number one problem this author sees is that entrepreneurs fail to recognize that differences between their motivations for starting businesses and the investors' motivation for funding them. Publicly available examples are used throughout this chapter to help explain the additional volatility that entrepreneurs face when using venture capital to fund their e-business start-up companies.

Beginning of the Dot-com Gold Rush

The mid-1990s marked the beginning of the dot-com gold rush, a period of time when investors eagerly supplied capital to any entrepreneur willing to brave the mysterious world of electronic commerce and build a business-to-consumer (B2C) company presence on the Internet. All would-be entrepreneurs needed were the words "online," "dot-com," or "electronic commerce" in their business plans to spark significant interest. Traditional concerns regarding the potential to generate revenues—typically measured by the existence of current sales—were temporarily suspended, as investors believed that the Internet held limitless potential and that existing business methods for sales held little predictive value for the success of an e-business. Essentially, as recently as 1993 the boundaries of this new frontier were completely unknown and every investment was a tremendous leap of faith. Yet many investors readily made this leap, believing that the payoff would be worth it.

Since 1993, we have learned a great deal about the Internet's role in business, including the fact that many of the keys to success for an e-business venture are the same as the keys to success for most traditional business ventures. "Just because you're part of the Net doesn't mean that the laws of economics have been repealed. You've got to have a real business," says Bill Reichart, the current President of Garage Technology Ventures (Aragon, 2000). The challenge is in finding the right balance between traditional and new business measures and methods. Many e-business ventures that were started in the mid-1990s continue to flourish, but many more ventures have come and gone. What we hear about most in the popular press are the spectacular failures of the dot-com companies, often dubbed "dot-bombs." In fact, whole Web sites have been dedicated to documenting these failures. Table 1 lists some of the failed companies' documentation sites, still up and running in September 2002. It is interesting to note

Table 1 Listing of Dot-com Failure Web Sites

HOOVERS	http://www.hoovers.com/news/detail/0,2417,11_3583,00.html
ITWATCHDOG	http://www.itwatchdog.com/NewsFeeds/DotcomDoom.html
PLANETPINKSLIP	http://www.planetpinkslip.com/
DOTCOMSCOOP	http://www.dotcomscoop.com
WEBMERGERS	http://www.webmergers.com/

that at one time many more documentation sites existed, but many of them ceased to exist as the economy slowed considerably during the third quarter of 2001 and continued to decline for several months afterward.

Studying past failure is an important part of increasing future likelihood of success, lest history repeat itself. Many articles have documented the explosion of e-business failures, with particular emphasis given to the myriad new but bad ideas that have been funded. However, that's not the entire story, or even the most interesting bit of the story. In this chapter, a brief discussion of other major reasons for dot-com failures precedes the discussion of keys to business plan success. Specifically, this chapter focuses on two additional key factors in numerous dot-com failures. First, the rapidly changing interests of investing firms are discussed. That is followed by a discussion of failed frontrunners with new and good business ideas that just lost their balance on the steep learning curve of new customer issues and/or business infrastructure requirements for e-business success.

Changing Interests of Investing Firms

Beginning in the mid-1990s, numerous investors eagerly sought electronic commerce ventures and were especially interested in funding B2C start-ups. The prospect of potentially reaching any consumer around the world without the cost of setting up physical shops seemed to hold unlimited profit potential. In fact, for a short period of time it seemed that the ultimate goal was to find and fund purely virtual companies, or companies completely independent of physical constraints. One of the first successful virtual companies is Tucows (http://www.tucows.com). Scott Swedorski, a young man who was very good at developing Winsock patches for Microsoft products, started Tucows in 1993. Mr. Swedorski developed a Web site and posted software patches as freeware. The Web site became so successful among software developers that advertisers rushed to place banner advertisements on the Web site. It is interesting to note that not only is Tucows one of the first successful virtual companies, it is also one of a very few companies that successfully used banner advertising as a sustainable business model. Indeed, Tucows is the exception, not the norm. Mr. Swedorski recognized a unique market need and fulfilled it, using the Internet.

In contrast, many new e-business start-ups have been driven more by a desire to identify new Internet markets rather than by an existing need. That is one valuable lesson from the Tucows example. Another lesson is that Tucows did not need venture capital funding because the company grew as a result of meeting a need in a new way. There is much less risk all around if an entrepreneur with a good idea can start small and grow the business slowly. Entrepreneurs who want to start big usually require venture capital. Invariably, this means that entrepreneurs will relinquish control of company management. It also means that overall the financial stakes are higher.

Two examples of start-up failures are briefly documented here to illustrate the related risks of starting big and using venture capital. Please note, however, that these examples are not intended to represent an indictment of VCs. Rather, it is a call to attention of the challenges of using venture capital that are generally overlooked.

Pets.com started off as a good idea—one-stop online shopping for pet supplies, from food to beds to routine medications. Pets.com was started late in the 1990s, a time when VCs were pushing managers to establish brand recognition at all costs. During this period, it was common for VCs to give large sums of money to start-up managers and tell them to spend aggressively on advertising campaigns designed to establish significant brand recognition as quickly as possible. During 1998 and 1999, the spectacular successes of a few firms with this strategy further encouraged other firms to push the limits on advertising campaigns. For example, exorbitant spending by the start-ups monster.com and hotjobs.com during the Superbowl in 1999 proved to be very successful. There was little reason to believe that it wouldn't work for other companies, such as Pets.com. Yet in the fiscal year ending September 30, 1999, Pets.com made $619,000 in revenues and spent $11,815,000 on marketing and sales. They were arguably successful in establishing brand recognition, as most people today would still recognize the Pets.com sock puppet, but the cost was too high. In the year 2000, they went public, but despite their brand recognition, consumers did not come in as quickly as as needed to build revenues. Consequently, they closed their doors the same year they went public.

The role of VCs in the failure of Pets.com and similar companies during the late 1990s is often overlooked in the retelling of the stories but was, in fact, key to those failures. Instead of fostering the vision of a sustainable business model, many VCs pushed artificial company growth purely through venture spending, with the goal of taking the company public as quickly as possible. Once such companies went through their initial public offering and public investors started to buy, the original VCs often sold their shares. Thus, it was traditional investors, not the VCs, who were left holding the losses. Traditional investors have learned the hard way that revenues matter, even for young companies. At the same time, for would-be entrepreneurs interested in seeking venture capital, such stories provide important lessons: Logistics and transportation issues probably would not have spun so far out of control if companies had started smaller, and brand recognition is not the only thing that matters.

It is important to keep in mind the focus of high risk investors such as VCs: They are willing to make higher risk investments with the hope of fast and significant payoffs. Unfortunately, this translates into an environment where some of those investors are willing to sacrifice the long-term future of a start-up company for their own profitability. Accordingly, it is very important that entrepreneurs check the track records of different investors before deciding whom to approach for funding.

VCs also played a major role in the failure of companies that never went public simply by shifting their interest to other types of start-ups before their previous start-up investments had stabilized. Streamline.com was an online grocery company that had a business model superior to most of their competitors' in terms of provisions for customer value and service. One of their primary problems was that they underestimated logistics and transportation costs, and they also found that competing with large grocery stores for top-quality products from food production companies was difficult. By the time that they began to learn from these issues, the VCs decided against further financing of the business, as their interests had shifted to other, newer types of emerging e-commerce companies.

The initial enthusiasm for B2C start-ups in the mid-1990s was replaced in the late 1990s with interest in the business-to-business (B2B) space. Investors began to realize that the profit potential in the B2B space was much greater because of larger transaction size and volume and greater customer reliability (businesses instead of individuals). Soon after B2B business plans became the favored choice for VCs, investors began to realize that a lot of work on Internet infrastructure was needed in order to truly exploit the Internet's potential. Thus, companies building software applications and other support mechanisms for strengthening the Internet's infrastructure replaced B2B investments as the favored business plans to invest in.

Within the span of 5 years or so, the interests of investors moved from B2C, to B2B, to Internet infrastructure business plans. This was taking place at the same time entrepreneurs were receiving large sums of money to build artificial growth. The result was that a number of companies with sound business plans grew too quickly to be sustainable on their own revenues, yet venture backing stopped, as VCs moved on to what they deemed to be more exciting categories of business plans. Streamline.com is but one example of a B2C company that started to show signs of progress as interest in financing shifted from B2C to what were considered potentially more lucrative ventures.

The moral of the story is that entrepreneurs looking for venture backing need to understand trends in business as well as what investments are favored by venture firms. Entrepreneurs willing to start smaller and grow more slowly can maintain control of the company's management and can move along the learning curve at a more reasonable pace. On the other hand, entrepreneurs with clear ideas for fulfilling significant and known market needs with electronic commerce businesses should genuinely consider venture capital as a funding option.

Confounding Organizational Issues

In the previous section, Streamline.com was mentioned as a firm that could not attract additional rounds of funding as they were learning how to manage their operations. Many online stores, such as Streamline.com and Amazon.com, started off believing they would compete directly with physical stores. Indeed, during the mid-1990s there was much discussion regarding anticipated changes in profits of physical stores and anticipated changes in consumer behavior. In the end, it became clear that most consumers were primarily interested in online shopping as an additional convenience rather than as an exclusive alternative to shopping in physical stores. This meant that online stores did not grow their market shares as quickly as they had hoped or anticipated.

Also, online stores underestimated their own dependence on companies with a physical presence, such as warehouses and distribution centers. Amazon.com is a perfect example of a company that ultimately was reorganized when managers recognized they needed the physical world more than they originally imagined (Vigoroso, 2002).

Brick-and-mortar companies making e-business investments struggled less with e-business investments, as they already had the infrastructure support they needed and they also had a clearer understanding of their customer base. In fact, many brick-and-mortar companies learned very quickly how to adopt best practices from the efforts of virtual companies and had the capital to implement their ideas with much less trouble than the virtual firms. Yet it is also interesting to note that many of these brick-and-mortar ventures into e-business lacked the innovation and forethought that could have resulted in truly innovative business practices. For example, Barnes and Noble developed their Web presence completely separately from their physical presence. Barnesandnoble.com had a completely different staff and completely different databases. Once the managers of the brick-and-mortar establishment felt that the market had stabilized and that the threat of market share loss from such companies as Amazon.com had dissipated, the managers of the brick-and-mortar Barnes and Noble decided to keep some shares in barnesandnoble.com but sell off the rest of virtual Barnes and Noble. Imagine, however, what innovations might have developed if, for example, the online databases had been fully integrated with the databases of sales from physical inventories? A much more complete understanding of customer spending habits on books could have been obtained.

Organizing the Keys to Success

Many of the keys to success for e-business ventures are simply the keys to success for any new business. The business world has learned a great deal by doing postmortem analyses of failed e-business start-ups. We have learned that customers want Web options as opposed to Web exclusivity in their lives. We have relearned that revenues count, relationships with suppliers are important, and a host of other traditional business values are still relevant. Many core considerations for building a successful company did not change as a result of the high tech investment boom and have not changed since the investment bust. Yet there are also some unique issues to tackle for e-business start-ups. The next four sections of this paper provide a layered view of methods for building a successful e-business plan, starting with the most general success considerations and building to the most specific issues. In particular, keys to success are analyzed as follows: considerations for any new business venture, considerations for any new company, special concerns when planning a new e-business venture for an existing brick-and-mortar firm, and special considerations when starting a new business that is primarily an e-business.

CONSIDERATION FOR ANY NEW BUSINESS VENTURE

A business venture may be anything, from an investment in a major new project by an existing company, to the start-up of a new company. Although new and established companies have some unique challenges and opportunities, there are also numerous issues common to both types of companies. Regardless of the size and sophistication of any given company, it is imperative to begin with three basic questions: what, who, and how much. More specifically, what is the venture goal or strategy, and how is the proposed business plan different from competitors' plans? Who will be designated as the key leaders for the venture, and approximately how much capital is needed to start the project or company? This section takes a closer look at each of these issues.

Identification of a Competitive Mission

The first thing angel investors or VCs will look at in a business plan is the executive summary. Specifically, they will be looking to see how the mission of a company is different from others they have seen. Investors will be looking for evidence that the company is proposing products or services that are in demand and that the organizational leaders in some way demonstrate innovation in their thinking about running the business. They will also be looking for evidence that the key managers have given adequate thought and consideration to the competitive environment. Investors want assurance that thorough analysis of the customers, competitors, and suppliers has been completed. Since the failure of so many e-business start-ups, investors have started scrutinizing mission statements very carefully and asking organizational leaders numerous difficult questions prior to making an investment. In a sense, this is an advantage to the proposed organizational leaders, as they should be forced to consider many possible business scenarios prior to actually starting their business.

Established brick-and-mortar companies investing in e-business operations, either B2B or B2C, should ask themselves the same difficult questions that VCs and bankers will ask entrepreneurs, to ensure that the proposed business plan fits well with the organization's existing missions and goals. However, brick-and-mortar companies often find themselves in a position of investing in e-business out of competitive necessity rather than out of competitive strategy. As a result, the e-business investments of the brick-and-mortar companies are often not leveraged to the extent they could be if they were more strategically linked to organizational goals and missions. Barnes and Noble is an example of lost opportunity when a major investment in e-business is never strategically linked to other core services offered by the company.

Key Employees

It is important to identify project leaders and key personnel as early as possible. These individuals need to provide strong leadership to the organization and to shape the company's future. Yet aspiring entrepreneurs often learn that finding a good business partner is about as difficult as finding a good marriage partner. First and foremost, entrepreneurs should find someone whose skills complement their skills. For example, a software engineer needs someone to be responsible for designing a marketing and sales plan. Also, entrepreneurs need partners who share their vision and, perhaps most important, are capable of helping fulfill that vision. Many would-be entrepreneurs are enthusiastic at the beginning of a start-up venture but do not know how, or do not have the wherewithal, to see the vision through to fruition. It is generally not a good idea to jump into a business venture with a relative stranger. A possible exception is if the partner candidate has a proven track record and enough time is spent with him or her to know that a good rapport exists. A proven track record for follow-through is of equal importance to rapport, however, and should not be overlooked simply because the person is likable. Anyone who can't execute your business plan effectively will quickly become less likable!

Ronald Grzywinski and Mary Houghton, cofounders of Shorebank Corp (http://www.sbk.com) note that effective management also requires a commitment to developing and creating opportunities for others to move through the ranks of the organization. They find that it is most difficult to find personnel with strong general management skills. Accordingly, they invest a great deal in building these skills in their existing personnel.

Identifying key personnel for e-business ventures of brick-and-mortar companies has its own challenges. Identifying whom to designate as the new managers for the e-business venture of a brick-and-mortar company involves deciding whether to move proven managers from the existing organization over to the e-business, or whether to hire new personnel from outside the company. Existing employees can ensure that an understanding of the current customer base is considered as decisions are made to develop the new venture. On the other hand, it may be possible to attract e-business veterans who can help the traditional management team with the e-business learning curve. It seems like an attractive option to hire a combination of insiders and outsiders to run new e-business ventures, but care must be taken to ensure that a culture clash does not get in the way of effective management.

In any event, it is most important that the strategic goals and missions of the e-business be clearly thought out and articulated at the very beginning, and that every subsequent decision be measured against these goals.

This will ensure that the scope of projects considered by the company stays reasonable and that the company does not take off in directions that cannot be subsequently supported by the core capabilities of the organization. Although this sounds like common sense, managers often get so caught up in either making a sale—any sale under any circumstances—that they forget to focus on fulfilling the strategic missions of the company. A clearly articulated and carefully considered mission that is continuously referenced by the management team can significantly reduce the likelihood of miscommunication within the organization and will help keep all personnel on the same strategic page.

Budgeting

Budgeting for any new business venture is not easy. E-business ventures, however, are particularly difficult to budget for, as the majority of capital used in start-up is for Web and other software development, network infrastructure, personnel, marketing, and other intangible items. Budgeting for intangibles is difficult because costs are difficult to estimate and the resulting financial benefits are difficult to calculate. Yet without adequate effort to capture estimated costs and benefits, financial control of a company is quickly lost. Oracle financial managers note that traditional business planning relies heavily on return on investment (ROI) as the key metric for measuring project effectiveness. However, to accurately measure e-business effectiveness, organizations are looking at multiple additional metrics, including changes in productivity, changes in cost of sales, and changes in the costs of customer acquisition and the gains in customer retention (Oracle, 2001). In any event, perhaps it is most important to remember that numbers associated with intangible assets are invariably based on numerous assumptions. Wherever financial calculations and projections are made pertaining to e-business, the assumptions should be readily available as well. Typical financial information that should be incorporated into the business plan include the following: sales projections, product demand, cost of sales/services, breakdown of fixed and variable costs, projected months to break even, projected months to positive cash flow, estimated capital burn rate, and a full set of pro forma financial statements.

While developing financial projections, it is also worthwhile to consider outsourcing some e-business activities. Even if managers ultimately decide to develop and manage everything in-house, consideration of outsourcing can help clarify cost and management issues. Common e-business activities that many companies have decided to outsource today include hosting of Web services, customer support applications, information security management, and advertising.

It is worthwhile mentioning that there are numerous budgeting tools available for e-business entrepreneurs on the Web. For example, at Entrepreneur.com, many forms are available for downloading, including forms to help managers estimate business start-up cash needs, business insurance planning worksheets, personal cash flow management statements, and development budget worksheets. There are also many stories in the online business periodicals documenting financial mistakes made in e-business start-ups. An article by Benoliel (2002) cites five of the most common financial mistakes made by new e-businesses.

Overstating Projections. Benoliel notes that like Enron, some companies may overstate expected revenues to deceive, but many companies overestimate earnings simply by being overly enthusiastic. In either case, realistic budgets and projections are the best choice for the long-term success of relationships between managers and investors.

Ignoring Immediate Budgetary Needs. Although some managers err by being overly optimistic, others will not ask for enough upfront capital to get started because they don’t want to scare off potential investors. Since the failure of so many e-businesses, managers have created more conservative budgets, believing that capital will be much more difficult to come by. Ultimately, however, there is still a lot of investment capital out there and investors are looking for good ideas to fund adequately for success. Again, best-estimate budgets are better than overly ambitious or overly conservative budgets.

Revenues = Positive Cash Flow. This is a simple accounting principle that many entrepreneurs without a business background can get into trouble over. Where possible, a conservative policy of delaying new purchases until after revenues are collected and existing bills are paid is a good business practice for young companies with minimal cash inflow.

Forgetting Taxes. Sales tax and employee withholdings are paid periodically instead of perpetually and accordingly may be temporarily forgotten as genuine expenses until it is time to pay them.

Mismanaging the Advertising Timeline

Advertising costs are often recorded as a percentage of sales in the same period. Yet advertising costs are actually incurred in the hope that they will lead to future sales. Failure to budget the appropriate items in a strategic time frame will underutilize finances needed to achieve sales goals and can lead to overspending in later months.

Security of Transactions

Eighty percent of the brick-and-mortar companies still not engaged in electronic commerce claim the primary reason for their trepidation is the inherent insecurity of the Internet medium. Thus, full consideration must be given to all security issues involved in the use of the Internet for e-business. Specifically, considerable thought should be given to security of information as it resides in company databases and as it is transmitted from point A to point B. Also, the confidentiality and privacy of customer information needs to be carefully considered. General security plans should be explained in business plans to the extent possible.

Many entrepreneurs consider the use of third-party Web site security providers as a means of ensuring transaction security and thus attracting new customers.

Specifically, some of the most frequently used companies for assuring information security, information privacy, or both include TRUSTe (http://www.truste.org), the Better Business Bureau (http://www.bbbonline.com), the AICPA (http://www.aicpa.org), and Verisign (http://www.versign.com). Major differences in offerings from these four organizations are briefly summarized here. For an in-depth analysis of the differences, the reader should consult the corresponding Web sites. Essentially, the Better Business Bureau is, and always has been, a consumer-oriented organization. As such, they focus primarily on assuring consumers that online vendors will respond appropriately to consumer complaints. Although this is the lowest level of assurance, it is the oldest company, so an entrepreneur may actually attract more average consumers with this seal than with lesser known seals that indicate stronger security. The TRUSTe seal assures consumers that an online company provides certain levels of protection of personal information in addition to promising to respond to customer complaints. Verisign, on the other hand, focuses on authentication of online merchants to assure that they are who they say they are, as well as on assurance that actual transactions are safely transmitted in an encrypted form to protect against data theft. Finally, the AICPA focuses on providing comprehensive transaction integrity assurance as well as on informing customers about specific business policies. AICPA services are far more comprehensive, and far more expensive, than other organizational offerings.

NEW START-UP CONSIDERATIONS

Anyone thinking about starting a new company needs to consider at least two issues, in addition to those discussed in the previous section. Specifically, would-be entrepreneurs of new companies need to consider how they will go about building brand recognition and customer loyalty, and how they will go about financing the start-up of their company. These are concerns for any new company, high tech or otherwise.

Building Brand and Customer Loyalty

It costs a company five times as much to acquire a new customer as it does to keep an existing one. This is good news for existing companies building e-businesses but rather daunting news for a start-up company. The fact that acquiring new customers is difficult has lured many e-business entrepreneurs into spectacular spending on advertising campaigns. As mentioned in the introductory section of this chapter, there was a brief period during the late 1990s when e-business managers focused on establishing brand recognition at the expense of all other business considerations. The fact that it worked for companies like monster.com fueled overspending by many e-business managers, who followed their lead. Another example is Drkooop.com. Prior to their demise, they spent very heavily on advertising. Specifically, they had deals with Creative Artists Agency, ABCNews.com, and three of Disney's GO Network sites to be their exclusive health content provider. In July 1999, drkoop.com signed a four-year, $89 million contract to feature its logo on AOL's portals—an amount that represented nearly 181% of the company's total available cash at the time. Although traffic increased dramatically, the price paid by Drkoop.com proved too high, just as it did for Pets.com. In December 2001, Drkoop.com filed for bankruptcy. Although we now have proof that focusing only on brand recognition is more often than not a fatal business concept, establishing brand recognition is still a critical success factor for start-ups. Balancing spending to establish brand with all other start-up costs when budgets for new companies are becoming quite lean is particularly challenging.

A good starting point for spending on brand is to identify a target market and build advertising campaigns specifically matched to interests of individuals in that target market. Media choices should be carefully selected based on target market habits. Even if a generous budget is set for advertising and marketing, a well thought out plan should drive spending on brand. Otherwise it is very easy for this expense to spin out of control. Sales and marketing personnel should be given liberal freedom to spend, but expenses should also be reconciled against plans, and personnel should be expected to justify expenditures on a periodic basis. Perhaps the most infamous example of why justification and reconciliation are so important is the story of the six Barclay bankers who ran up a $62,700 dinner tab (mostly for drinks) at a fine restaurant in London and then tried to pass it off as a client expense. Reconciliation and justification efforts resulted in five of the six bankers getting fired. Thankfully, most entrepreneurs do not need to worry about their sales personnel spending $62,000 on one dinner, but if management makes it clear early on that all employees are accountable for the money they spend, they will think twice before treating themselves to airline upgrades, expensive hotel rooms, valet parking, and other unnecessary luxuries commonly expensed when no one is watching.

Building customer loyalty is related to building brand recognition in that having loyal customers is one of the most effective means of establishing brand. Yet customers are usually loyal to companies that provide unique products and services, or that provide products and services in a consistently reliable fashion or at consistently lower prices. Accordingly, building customer loyalty is particularly challenging for B2C start-up companies that aim to sell commodity-type goods over the Internet. A broad definition of commodity is used here: It means that products manufactured with the same specifications are not different from each other. In other words, two items that are configured in exactly the same way are, for all intents and purposes, interchangeable. Books, CDs, computers, and automobiles are examples.

Financing Options

As with many other aspects of business, favored financing options for start-up companies have in a sense come full circle. Before the e-business boom, many new entrepreneurs built their businesses slowly, by starting with one or two clients, and then applying revenues from those sales either directly to investments in a new company or as collateral for a traditional bank loan.

The dream of early e-business entrepreneurs, however, was to grow their company quickly to a substantial size. What we have learned from the growing pains experienced by many of those firms is that slow and steady has definite merits over fast and fatal! For example, Donna LaVoie, entrepreneur, launched her corporate communications company in the fall of 2001 (Hendricks, 2002). She took a one-third prepayment of services to be rendered for her first client and used that to hire her first employee.

This return to old-fashioned values may not be entirely by choice for many entrepreneurs, however. Results from a recent survey by the National Association of Manufacturers show that more than a third of respondents find it harder to get loans from their banks now than it was in early 2001, whereas results from another survey, by the National Venture Capital Association, reveal that the number of companies receiving venture money fell from 6,245 in 2000 to 3,736 in 2001 (Henricks, 2002).

Yet there is also evidence that capital is still readily available for well-presented business plans. Enter the term venture capital into any good search engine, such as Google, and a very large number of venture capital firm sites will be returned. A few especially good sites to begin with include http://www.Start-upjournal.com, http://www.businessplans.org, and http://www.garage.com. Before approaching VCs, however, entrepreneurs interested in external funding should understand a little bit about the basic categories of financing options. New start-ups with no initial capital and no customers will usually start by looking for "angel" investors. Angel investors are individuals or companies willing to take bigger risks with their capital than VCs are, but they will also expect bigger company percentages than traditional VCs, generally in the form of a larger share of the start-up company. Angels generally expect to contribute anywhere from $100,000 to $1,000,000,000 to help get a company started. VCs today will come in a bit later and large venture capital companies will often want a seat on the company's board of directors to help protect their investments. Another investment option is to approach companies that specialize in providing management services in addition to financial backing. These companies are called incubators.

This chapter has described the numerous advantages and disadvantages both to self-financing and to external financing of operations, with particular emphasis on the pros and cons of using venture capital. These issues need to be carefully considered before financing options are pursued in earnest.

SPECIAL CONSIDERATIONS FOR E-VENTURES OF BRICK-AND-MORTAR FIRMS

Brick-and-mortar firms need to consider the issues discussed in sections one and two of this chapter, as well as the issues unique to their environment, which are described in this section. Specifically, brick-and-mortar companies have unique customer and operational management issues to sort out.

Potential for Channel Conflict/Channel Complement

Channel conflict occurs when a company cannibalizes its own sales with other operations. For example, eSchwab cannibalizes some of Schwab's traditional services and fees, but other online brokerage services were capturing part of their market share, so they felt they did not have a choice. At the same time, Schwab did a good job of building synergies between their online and offline service offerings, ultimately resulting in channel complements that helped build customer loyalty.

In retail, many companies have developed an online presence, believing it is a competitive necessity. This often leads to a situation where little is truly gained by the online presence, as little strategic intellectual capital is invested. One company that has done a great job of building channel complements between its online and offline business is Talbots. Talbots was one of the first retail companies to embrace the idea that customers want the convenience of being one customer to the entire company. Accordingly, you can buy a product at Talbots.com and return it to a brick-and-mortar Talbots store with no hassle. This has done a great deal to boost their sales overall.

Most businesses lose about 25% of their customers annually. If brick-and-mortar companies viewed e-business investments as a means of keeping their existing customer base happy, they would likely be more profitable overall, by reducing the number of customers lost.

Operational Management Issues

The first major management decision that brick-and-mortar companies need to make is whether to manage the online operations completely separately from the offline business. The temptation is great to manage them separately because it is far simpler and more familiar. Yet separate management increases the likelihood that the two different channels will end up competing for the same customers. Separation also increases the risk that the online operations will not stay aligned with the corporate business missions and goals. There are tremendous opportunities for creative thinking about channel management and building customer loyalty if a company decides to operate the online and offline operations synergistically. Yet commitment to building synergies is likely to be very expensive, as team management issues and data architecture issues may require significant upfront capital outlays.

In any event, strong management leadership and a commitment from top management is a major key to the success of e-ventures made by brick-and-mortar companies.

One advantage that brick-and-mortar companies have over virtual companies is that they are more likely to have good business policies and practices in place for measuring success of investments and for evaluating management performance. Constant feedback on performance is key for a new venture. Another advantage brick-and-mortar companies have is that they start with an existing customer base. Identifying creative means for leveraging this base can be the difference between success and failure.

SPECIAL CONSIDERATIONS FOR NEW E-BUSINESSES

Finally, there are a couple of special considerations for new e-business entrepreneurs. Specifically, new e-businesses must give special consideration to bandwidth and security, in addition to the issues discussed in the first three sections of this chapter.

Bandwidth

A company that is completely dependent on the Internet for sales must ensure that customers have minimal wait time for pages and images to appear on screen. This is a life and death issue for start-up companies. It is important to select top-quality Internet service providers and to have highly scalable software solutions. The user interface should be attractive and helpful but should not be unnecessarily bandwidth intensive.

Additional Security Issues for New E-businesses

It is fairly well known by now that companies with an online presence attract computer criminals. The latest annual security survey from the Computer Security Institute revealed a number of shocking statistics, including the following findings (http://www.gocsi.com):

Ninety percent of respondents detected computer security breaches within the past 12 months.

Eighty percent acknowledged financial losses due to computer breaches.

For the fifth year in a row, more respondents (74%) cited their Internet connection instead of their internal systems as a frequent point of attack.

Forty percent detected system penetration from the outside.

Forty percent detected denial of service attacks.

Although all companies with an online presence need to ensure the best security possible for online transactions, for a company that is completely dependent on online customers, one security breach is potentially fatal to the entire company. Thus, new e-business entrepreneurs need to ensure that they budget adequately for security of financial and personal customer data, security of financial and personal employee data, appropriate internal and remote access controls, control over portable and handheld computing devices, and transaction security mechanisms and policies.

PREPARING THE BUSINESS PLAN

In this final section, all the elements described in previous sections are pulled together. The key elements of a modern business plan are described and an outline of essential e-business plan elements is provided. Writing a business plan for an e-business is not much different from writing one for any other business now. If anything, what has changed since the investment in, and subsequent failure of, so many Internet companies is that investors have become a lot more savvy about their investments and scrutinize business plans greater than ever. Investors in new companies like to take calculated risks and are usually looking for new and exciting ideas to invest in. In the business plan, the entrepreneur must communicate effectively yet simply the innovation in his or her business ideas. It is critical that the business plan include a reasonable explanation of the value of the company's ideas.

Business plans are generally shorter than they used to be, because investors expect either to be involved in company decision making or to keep close contact with the start-up company's management. First and foremost, a business plan should begin with a good executive summary. If the executive summary doesn't hook potential investors, they will not read the rest of the business plan. Following the executive summary, a business plan should have at least four major sections: a detailed description of the business, a marketing plan, a financial plan, and a general management plan. There are a multitude of Web-based resources for writing a business plan (Table 2) and there are a number of possible additions to a business plan beyond the basic four elements described above. The model in the Appendix provides a reasonably comprehensive description of the business plan sections most commonly found in an e-business plan.

Ultimately, whether a business plan is read depends on whether it sparks interest with the reader. Accordingly, entrepreneurs should get to the point of uniqueness quickly, emphasize why their team can fulfill the concept better than any other company, and put forth reasonable financial goals for getting the job done. If these key elements are in place, the end result will likely be successful funding of the business proposal.

Table 2 Sources for Writing a Business Plan

BUSINESS 2.0	http://www.business2.com/webguide/0,1660,19799,FF.html
INC	http://www.inc.com/guides/start_biz/directory.html
DELOITTE	http://www.deloitte.com/vc/0,1639,sid%253D2304%2526cid%253D9021,00.html
SBA	http://www.sba.gov/starting/indexbusplans.html
BUSINESS PLANS SOFTWARE	http://www.businessplansoftware.org
GOOGLE DIRECTORY	http://directory.google.com/Top/Business/Small_Business/Resources/
WSJ's START-UP JOURNAL	http://www.Start-upjournal.com
BUSINESS PLANS	http://www.businessplans.org

APPENDIX

The following is an excerpt from Professor Dennis Galletta's Web site on Business Plans (http://www.pitt.edu/~galletta/iplan.html). It lists the sections of an e-commerce business plan.

Executive Summary: This section must concisely communicate the basics of an entire business plan. Keep in mind that your reader may be unfamiliar with the Internet and its tremendous potential.

Business Description: In this section, discuss your firm's product or service, along with information about the industry. Describe how your product and the Internet fit together or complement each other.

Marketing Plan: Discuss your target market, identify competitors, describe product advertising, explain product pricing, and discuss delivery and payment mechanisms.

Customers: Define who your customers are and how many of them exist on the Internet. An analysis of the customer base should not be a casual guess.

Competitors: Use Internet search engines to look for known competitors or products similar to yours. Be sure to use several search engines, because each uses different search techniques. All major direct competitors should be found and analyzed in your plan.

Remember, readers of your business plan will be very interested in knowing how you are going to beat the competition.

Advertising: Describe how you are going to tell the Internet community about your product or service. Designing beautiful Web pages is only a first step. You must also get the word out about your Web site. Some tips: Detail a plan to add your Web address to the databases of search engines, add it to the bottom of all of your e-mail messages, and perhaps create physical novelties for local customers.

Pricing: How are you setting prices for your products or services? If your product is intangible information delivered over the Internet, you should try to create some sort of pricing model to justify your prices. You could start by researching what others are charging for similar products.

Delivery and Payment: How are you going to deliver your product and get paid? E-mail alone is not secure. Consider encryption techniques and online payment services.

Research and Development: The technical aspects of your company, this addresses where the company is now andd the R&D efforts that will be required to bring it to completion; it will also forecast how much the company will cost. Since the Internet is continually developing, you should also address continuing plans for R&D.

Operations and Manufacturing: Discuss the major aspects of the business, including daily operations and physical location. Also, what equipment will your business require? Will you be using your own Web server, or will you be contracting with another company? Who will be your employees—will you hire staff with knowledge of the Internet or will you train them in-house? Be sure to include cost information.

Management: Address who will be running the business and their expertise. Because the business centers around the Internet, be sure to discuss the management team's level of Internet expertise and where they gained it. Also, describe your role in the business.

Risks: Define the major risks facing the proposed business. In addition to such regular business risks as downward industry trends, cost overruns, and the unexpected entry of new competitors, also include risks specific to the Internet. For example, be sure to address the issues of computer viruses, hacker intrusions, and unfavorable new policies or legislation.

Financial: Include all pertinent financial statements. Potential investors will pay close attention to this area, because it is a forecast of profitability. Remember to highlight the low expenses associated with operating on the Internet compared to those of other business.

Timeline: Lay out the steps it will take to make your proposal a reality. When developing this schedule, it might be helpful to talk to other Internet businesses to get an idea of how long it took to establish their Internet presence.

GLOSSARY

Angel investor Early investors in new Start-up companies who usually are willing to accept even higher risks than venture capitalist but in exchange for anticipated larger returns, generally in the form of a larger share of the Start-up company for their investment than later venture capitalists would expect to receive.

Brick-and-click companies Companies with both traditional operations and e-business operations. Originally, brick-and-click was used to describe brick-and-mortar companies that built an e-business presence. Now many e-businesses are also building traditional operations.

Brick-and-mortar companies Companies without any e-business presence or to traditional companies whose operations are completely dependent on physical buildings and other assets and physical business infrastructures.

Business-to-business (B2B) Internet age term referring to the online exchange of products, services or information between two or more business entities.

Business-to-consumer (B2C) Internet age term referring to the online sale of products, services or information to consumers by businesses.

Channel conflict When a company cannibalizes their own sales with other operations.

Dot bomb A failed dot-com company.

Dot-com A dot-com is any Web site intended for business use and, in some usages, it's a term for any kind of Web site. The term is popular in news stories about how the business world is transforming itself to meet the opportunities and competitive challenges posed by the Internet and the World Wide Web (definition taken from www.whatis.com).

E-business venture Any major investment in an e-business initiative ranging from investment in a new Web-enabled transaction processing system for a brick-and-mortar company to a new virtual company built around a set of e-business missions and goals.

Incubators Companies that specialize in providing management services in addition to financial backing

Venture capitalist (VC) Investors, usually in smaller private companies, who are looking for larger than average returns on their investments. VCs may be in private independent firms, in subsidiaries or affiliates of corporations, or government supported agencies.

Venture funding Investment funds received from a venture capitalist.

Virtual company A company with primary operations that are independent of other brick and mortar companies.

CROSS REFERENCES

See *Business-to-Business (B2B) Electronic Commerce; Business-to-Business (B2B) Internet Business Models; Business-to-Consumer (B2C) Internet Business Models; Click-and-Brick Electronic Commerce; Collaborative Commerce (C-commerce); Consumer-Oriented Electronic Commerce; Customer Relationship Management on the Web; E-marketplaces; Marketing Plans for E-commerce Projects Electronic Commerce and Electronic Business; Mobile Commerce.*

REFERENCES

Aragon, L. (2000). *VC P.S.: Ten myths and realities of VC.* Retrieved February 13, 2003, from http://www.circlenk.com/10-Myths-vc-vcps061400.htm

Balu, R. (2000). *Starting your start-up, fast company.* Retrieved February 13, 2003, from http://www.fastcompany.com/online/31/one.html

Benoliel, I. (2002). *Avoid these errors to avoid financial nightmares.* Retrieved February 13, 2003, from http://www.Entrepreneur.com/article/0,4621,297437,00.html

Healy, B. (2002). *Tracking the incredible shrinking venture funds.* Retrieved February 13, 2003, from http://digitalmass.boston.com/news/globe_tech/venture_capital/2002/0729.html

Henricks, M. (2002). *Consider the benefits of funding alternatives.* Retrieved February 13, 2003, from http://www.Start-upjournal.com/financing/trends/20020501-henricks.html

International Council of Shopping Centers White Paper. *The marketing of a net company.* Retrieved February 13, 2003, from http://www.icsc.org/srch/rsrch/wp/ecommerce/marketingofanetcomp.html

Johnson, A. (2000). *Special report: Console makers face a brand-new game.* Retrieved October 29, 2002, from http://www.upside.com/texis/mvm/story?id=39aea60e0

Oracle White Paper (2001). *Essentials for a winning e-business plan.* Retrieved February 13, 2003, from http://www.oracle.com/consulting/offerings/strategy/epswp.pdf

Vigoroso, M. (2002). *And the e-commerce gold medal goes to. . . .* Retrieved February 13, 2003, from http://www.ecommercetimes.com/perl/story/16387.html

Business-to-Business (B2B) Electronic Commerce

Julian J. Ray, *Western New England College*

INTRODUCTION

The focus of business-to-business e-commerce (e-B2B) is on the coordination and automation of interorganizational processes via electronic means. E-B2B is conducted between business organizations and is a subset of all e-commerce activity, which includes, among others, business-to-consumer (B2C), business-to-government (B2G), and consumer-to-consumer (C2C) activities.

Before an in-depth analysis of e-B2B, it is worthwhile to note that many authors today identify e-commerce as one component of a more general form of electronic business termed e-business. E-business is generally used to identify the wider context of process automation that can occur between businesses and includes automating the provision and exchange of services, support, knowledge transfer, and other aspects of business interaction that do not necessarily result in a buy or sell transaction being executed. Within this document the terms e-commerce and e-business are used interchangeably and refer in either case to the wider context defined above.

E-B2B is a major driving force for the Internet economy. According to research from the Gartner Group (http://www.gartner.com), B2B is growing at a rate of 100–200% per year. With an estimated $430 billion in 2000, the total value of e-B2B activity was initially expected to exceed $7 trillion by 2004 with the largest shares in North America ($2.8 trillion), Europe ($2.3 trillion), and Asia ($900 billion), and 24% of all e-B2B transactions are expected to be performed electronically by 2003 (SBA, 2000). However, the recent slowdown in the global economy has caused a revision of these initial predictions to reflect changing economic conditions. Current estimates for 2004 global B2B revenues have been reduced to $5.9 trillion with a growth rate of 50–100% per year (Gartner Group, 2001). Even with the slowdown, the rates of adoption are significant and on track to exceed $8.5 trillion in 2005.

FOUNDATIONS OF B2B E-COMMERCE

Sustaining a high rate of adoption and growth in e-B2B activity requires several key components: the innovative application of technology, a ubiquitous communication medium well suited for the transmission of e-B2B-related documents and information, a business and regulatory environment suitable for sustaining the process, and last but not least a set of motivating forces prompting organizations to adopt interbusiness automation as part of their day-to-day operations.

Innovations in Technology

The technological foundations of e-B2B have been developing over the past 30 years and reflect innovations in telecommunications, computer communications protocols and methods, and intraorganizational business process management. During the 1970s, electronic funds transfer (EFT) and electronic data interchange (EDI) were nascent communication standards developed for business exchanges over expensive private computer networks. The costs of participating in these initial systems were high. EDI systems, for example, require specialized data processing equipment and dedicated telecommunications facilities, prohibiting all but the largest businesses to participate. However, the potential benefits realized from implementing these early electronic B2B systems included greatly reduced labor costs and processing time as well as vast increases in accuracy.

Private communications networks requiring dedicated telecommunications facilities, initially the only option for e-business, are still in use today and provide a secure medium for exchanging electronic data. A value-added network (VAN) is a type of semiprivate communications network usually managed by a third-party provider that allows groups of companies to share business information using commonly agreed-upon standards for transmitting and formatting the data. VANs have traditionally been associated with EDI and provide an alternative to more costly private communications facilities.

More recently e-business has adopted the public Internet as its basic communication medium. The Internet provides the ubiquitous connectivity and high data rates necessary to support modern e-business. The significantly reduced costs of operation and costs of entry of the Internet with respect to other communication systems such as traditional VANs or private lines makes the adoption of an e-business model more attractive and sustainable to smaller organizations.

One of the key technologies allowing the Internet to support e-business is the virtual private network (VPN). VPNs are a form of protected communication using a combination of encryption methods and specialized data transmission that allows companies to use the public Internet as if it were a private system. VPNs also enable development of extended private networks (extranets), which are shared with business partners. Today VPNs largely replace the role played by early value-added networks with a cost-effective alternative allowing companies to freely share private business information over the public Internet.

Processing transactions electronically can significantly reduce the cost and time taken to manage business information. Computer servers today can handle thousands of transactions per second and significantly help reduce the costs and risks associated with sending and recording a business transaction. One of the greatest current interoperability challenges is the development of common methods and protocols to enable semantic translation. To date there are hundreds of protocols and standards that govern almost every aspect of using the Internet to transmit and translate business data. Extensible markup language (XML) is an example of a data-format technology that can be understood and processed by a wide variety of computer languages and systems.

These new and emerging standards build on earlier protocols such as EFT and EDI. When coupled with system-independent data communication protocols such as the Internet's TCP/IP, these initial data translation protocols enable business data to literally transcend organizational and technological boundaries. For example, while EDI documents can still be sent over private lines, today it is also possible to complete the same transactions using the Internet and XML at a fraction of the cost of traditional EDI. Companies who invested heavily in the early technologies can maintain their original investment and still participate in modern e-business initiatives while companies who could not afford technologies such as EDI are now able to engage in intraorganizational data exchange.

Innovations in Business Processes

Greenstein, O'Leary, Ray, and Vasarhelyi (in press), in a discussion on the electronization of business, identify the surge of e-B2B activity as a progression of an evolving process that has been evident throughout the industrial revolution. The authors attribute the rise of e-business activity as a response to dramatic improvements in both information processing and telecommunications systems, which have facilitated revolutionary change in organizational business processes. This most recent of technological revolutions centered around the Internet provides businesses with the ability to automatically place orders, send quotes, share product designs, communicate sales projections, and collect payment using electronic means, providing the potential to reduce operating costs, shorten time-to-market for new products, and receive payment faster than ever before. Moreover, businesses willing to commit to transitioning to the new digital economy have the potential to partner with other similarly minded businesses creating tightly integrated supply chains, remove intermediaries, and develop networks of loyal customers and trading partners who, research shows, collectively outperform their competitors.

Recent changes in the banking industry provide an example for understanding the potential improvements in business processes that can be achieved using electronic automation. The U.S. Federal Reserve along with 6 other central banks and 66 major financial firms recently implemented a system called the Continuous Link Settlement System. This system is designed to automate the trading and settlement of foreign currency exchanges. Traditionally, when performing a trade, two banks agree on an exchange rate and wire the money to each other. This is a process that can take two to three days during which time events can change the value of a country's currency, ultimately affecting the exchange rate. In extreme cases banks have gone bankrupt in this period of time (Colkin, 2002). Banks trade on the order of $300 trillion each day, so the ability to perform trades in real time can reduce the amount of lost interest. Accordingly, the risk associated with transaction latency justifies the $300 million investment in technology that the Continuous Link Settlement costs.

MOTIVATION FOR B2B E-COMMERCE

There are a number of reasons why a company might decide to invest in developing and implementing an e-B2B strategy. Established companies, faced with rising costs and increased competition, often have a set of motivating factors and goals different than that of newer companies. Newer companies are often more technologically agile and can readily take advantage of newer e-business strategies and products.

Benefits for Established Companies

E-B2B strategies have been proven to cut operations cost for companies and to reduce cycle time during product development. We provide examples from Federal Express and General Motors later in this text to illustrate how different e-B2B strategies can lead to cost reduction and an increased ability to compete by reducing the time required to design and develop new products. The ability of a company to automate its supply chain is another significant advantage of e-B2B. Reducing costs and production times are only a few of the types of benefits that can be realized by e-B2B companies. Other benefits are associated with increased visibility over the supply chain allowing for better planning and inventory management, strengthened relationships with business partners, and the ability to integrate with new business partners in a variety of interesting and innovative ways using mechanisms such as online exchanges and auctions.

Norris, Hurley, Hartley, Dunleavy, and Balls (2000) identify three stages in the adoption of an e-B2B business model by established companies. The early focus of a company's e-B2B activity are on increasing the efficiency of the sales and/or purchasing processes while minimizing the disruption to organizational culture and business processes. The second stage of adoption is usually on improving business processes by using electronic information technologies to integrate the supply chain and streamline the process of conducting business. This stage is often aimed at reducing costs and increasing the effectiveness of operations beyond sales and purchasing. The last stage in the conversion of a company to a mature e-business model is the development of strategic, tightly coupled relationships between companies to realize the mutual benefits and joint rewards of an optimized supply chain. General Motors (http://www.generalmotors.com) is an example of a traditional company that is rapidly evolving into an e-B2B company. General Motors has adopted a multipronged approach in its transition to an e-business using small pilot projects to test different e-B2B strategies in a drive to increase its competitive edge and reduce costs. Among GM's e-B2B initiatives are joint product design, dealer portals, and online auctions (Slater, 2002).

E-B2B can be used to reduce a company's risk such as the financial risks associated with completing monetary transactions in a timely manner as demonstrated by the Continuous Link Settlement System. E-business can also be an effective tool in reducing the risk associated with performing business under changing economic conditions. Such risk can be mitigated by developing economic as well as digital ties to existing business partners and providing a framework wherein new business partners can be introduced and incorporated into the supply chain.

New Products, New Services, and New Companies

E-business strategies may involve creating and introducing new products and services within existing companies as well as providing the basis for the creation of entirely new e-business. Online exchanges, for example, have emerged as a multibillion dollar industry virtually overnight as a result of the desire for increased speed and expediency of interorganizational electronic business transactions.

Proprietary supply-chain integration is another area with significant e-B2B activity. However, even with these proprietary efforts, companies acting as intermediaries may provide a large part of the technology and services involved in supply-chain operations. Examples of such companies include Manugistics (http://www.manugistics.com), which has successfully adapted its more traditional technologies and services to operate within the emerging e-business environment across a broad array of industries, as well as brand new technology-savvy companies such as Elogex (http://www.elogex.com) with niche strategies of developing technology and services specifically for a narrow industry focus: consumer goods, food and grocery in this case.

For small to mid-sized organizations, the barriers to participating in the e-business arena have been removed. The financial overhead associated with implementing early e-business systems such as EDI is significant and out of reach of all but the largest corporations. The rapid adoption of the Internet as a ubiquitous communication medium along with the reduced cost of computer systems and software over the last decade has allowed all types of companies to develop e-business strategies providing increasing levels of competition, innovation, and access to a global pool of potential business partners. In fact, the inverse relationship between size and agility gives small to mid-size firms some competitive advantages.

The dot-com boom of the late 1990s, which was so visible in the B2C sector with companies such as Pets.Com and WebVan, carried over into the B2B sector. During this time large numbers of highly innovative, technology-savvy Internet-focused companies became immersed in all areas of online business-to-business activity. By 2000 there were an estimated 2,500 online B2B exchanges serving industries such as electronics, health care, chemicals, machinery, food, agriculture, construction, metals, printing, and medical lab equipment (Wichmann, 2002). Like their B2C counterparts, however, many of the B2B dot-com companies have failed as a result of weak business models and misguided, overzealous venture capital. Now less than 150 of these exchanges are left.

CLASSIFICATION OF B2B STRATEGIES

E-business strategies are usually classified based on the nature of the interaction taking place between the business partners. Four strategies that may be classified using this approach are e-selling, e-procurement, e-markets, and e-collaboration. Within each of these strategies we can further identify the vertical or horizontal industry focus of the participation and whether the participation involves the use of intermediaries.

Companies with a vertical focus usually operate within a single industry such as the automotive, chemical, energy, or food-retailing industries. Alternatively, companies with horizontal focus provide products and/or services across a wide range of industries. Office supplies and computers are examples of products horizontally focused.

Intermediaries in e-B2B are most often associated with e-markets and electronic auctions where one company, the intermediary, acts as a broker allowing other companies to buy and sell products and services by collaborating using the infrastructure provided by the intermediary.

E-selling

E-selling is concerned with the direct sale of products and services to other businesses using automated means. The business model is similar to the B2C direct-sale model but differs in that B2B interaction requires prenegotiation of prices, catalog items, and authorized users. Turban, King, Lee, Warkentin, and Chung (2002) identify two major methods for e-selling: electronic catalogs and electronic auctions. Electronic catalogs can be customized in both content and price for specific businesses and can be coupled with customer relationship management software such as SAP's Internet sales solution (http://www.sap.com) to provide personalized content delivery. There are many examples where companies provide a direct sales channel to other businesses using the Internet: Dell Computers (http://www.dell.com) and Staples (http://www.staples.com) are examples of companies that allow businesses to set up accounts through the Web and customize the content available for online purchase. Cisco is another example of a company that has a very successful direct-sale approach using the Internet (http://www.cisco.com). Using Internet based pricing and configuration tools, Cisco receives 98% of their orders online, reducing operating costs by 17.5% and lead-times by at least 50% down to 2–3 days (Turban et al., 2002).

Electronic auctions are sites on the World Wide Web where surplus inventory is auctioned by a business for quick disposal. Auctions allow companies to save time and reduce the cost of disposing of surplus inventory, allowing the company to obtain a better price. Ingram Micro, for example, receives on average 60% of an item's retail cost by selling surplus inventory through an online auction. Prior to using the auction Ingram Micro recovered 10–25% of the price using traditional liquidation brokers (Schneider, 2002). Some companies create and manage their own auction sites such as CompUSA. CompUSA's Auction Site (http://www.compusaauctions.com) is designed to allow small and mid-sized companies to bid on its inventory. Alternatively, companies can use existing general purpose or industry-specialized auction sites such as eBay (http://www.ebay.com) or ChemConnect (http://www.chemconnect.com). Both Sun Microsystems and IBM have successfully adopted eBay as an auction site for surplus computer hardware, software, training courses, and refurbished equipment. eBay in this case is operating as a horizontally aligned intermediary between Sun Microsystems and the companies or individuals bidding on Sun's products. ChemConnect is an example of a specialized third-party auction site designed to manage the selling and buying of chemical and plastic products using an auction format. Unlike eBay, which is horizontally focused, ChemConnect has a vertical focus and concentrates on selling specifically to the chemical industry.

E-procurement

E-procurement is concerned with the automated purchasing of goods and services from suppliers. These applications are designed to facilitate the exchange of electronic information between trading partners by integrating a buyer's purchasing process with a seller's order entry process (Davis & Benamati, 2002). E-procurement systems can be either buy- or sell-side applications. Buy-side applications reside on the buyer's systems and control access to the ordering process, the authorization of the order, and possibly the list of trading partners allowed to receive the order. Sell-side applications reside on supplier's systems and allow authorized buyers to access and place orders directly on the system. Sell-side systems are often implemented as extranets where access to the electronic catalogs and pricing information is carefully controlled.

According to the Aberdeen Group, e-procurement is one of the main areas of e-B2B that is "delivering rapid and quantifiable results." They estimate that 80–90% of companies plan to use online procurement systems by 2003 (Aberdeen Group, 2001b). The sorts of products purchased by e-procurement systems are generally limited and of relatively low value. Strategis (http://www.strategis.gc.ca), an online branch of Industry Canada, report that 42% of e-procurement purchases by businesses in Canada in 2001 were for office supplies, furniture, and office equipment, followed by IT hardware and software (29% of total) and travel services (15% of total) (Strategis, 2002). Similarly, Microsoft reports that 70% of their annual 400,000 procurement purchases are for requisitions less than $1,000 (Microsoft, 2000).

Early e-procurement implementations relied on fax, e-mail delivery, EDI, or real-time system processing to transfer data among trading partners. Newer e-procurement systems use Internet technologies to manage the ordering process. Recent surveys indicate that most commercial solutions now use XML and the Internet to share procurement data and provide access to online market places as part of the offering (Aberdeen Group, 2001b).

E-procurement systems can significantly benefit companies in a variety of ways. In 1999 Federal Express identified e-procurement as a key strategy in reducing costs and saving time in the purchase order process. FedEx purchased a B2B commerce platform from Ariba Inc. (http://www.ariba.com), which was implemented within a month and reportedly returned a positive ROI within three months. The new system manages about 20% of FedEx's 25,000 annual requisitions and has reduced purchasing cycle times from 20 to 70%, depending on the type of items purchased. The purchase of new PCs now takes 2 days rather than the 17–19 days it took using a traditional paper-based approach. FedEx also managed to reduce the purchasing department staff by half, allowing these extraneous staff to be reassigned (Aberdeen Group, 2001a). Microsoft reports similar success with MS Market, a desktop procurement system designed to run from a Web browser over the company's corporate intranet. MS Market is deployed to 55 Microsoft locations in 48 countries and saves the company an estimated $7.5m annually (Microsoft, 2000). In the United States, Microsoft use MS Market to order almost 100% of the company's requisitions at an average cost of $5 per requisition.

E-collaboration

E-collaboration is a term broadly used to describe any form of shared e-business activity in which companies collaborate together with the goal of providing a mutually more efficient business environment. E-collaboration can take many forms such as supply-chain integration, the joint design and development of products, joint demand forecasting, and joint planning. General Motors, for example, shares specialized engineering software and data files with its suppliers in a drive to reduce product development time. As part of this joint-product design strategy, General Motors partially or fully underwrites the cost of the software licenses for some of its suppliers in order to standardize the design platform across the supply chain. By standardizing and sharing design tools, General Motors reduced the design-to-production cycle for new products by 50% to just 18 months (Slater, 2002).

Collaborative Planning Forecasting and Replenishment (CPFR) is a collaborative approach that allows suppliers and sellers to jointly forecast demand for products in an attempt to better coordinate value chain processes such as restocking, managing exceptions, and monitoring effectiveness. The major idea underlying CPFR is that better forecasting between trading partners will lead to improved business processes, which in turn will reduce costs and improve customer service and inventories. Wal-Mart's RetailLink is an example of a CPFR system that operates over the Internet using EDI documents to share information. RetailLink allows Wal-Mart's suppliers to receive detailed information about store sales, inventory, effects of markdowns on sales, and other operational information that allows the suppliers to effectively manage their inventory at the individual Wal-Mart stores.

CPFR and similar types of supply-chain integration can result in numerous benefits for companies interested in collaborating with business partners. After interviewing 81 companies, Deloitte & Touche concluded that companies that "collaborate extensively" with their supply-chain partners while focusing heavily on customer loyalty and retention are almost twice as profitable as companies that are "below average" in the areas of supply-chain collaboration. Deloitte & Touche also found that companies who understand customer loyalty are much more likely to exceed their goals on shareholder returns, return on assets, and sales growth, and 12% of the companies studied were in this group (Rich, 2001).

E-markets

E-markets provide third-party integration services by supplying online applications that allow organizations to exchange goods and services using a common technology platform. In essence, these electronic marketplaces are designed to bring suppliers and buyers together in a common forum with a common technology platform. E-markets are usually vertically aligned for industries such as energy or automotive but there are also horizontally aligned e-markets that service industries such as office supplies and information technology. Many e-markets are independent exchanges managed by a company that is neither a buyer nor a seller in the marketplace but provides third-party services that allow other businesses to collaborate through them. Alternatively, e-markets can be created and managed by a company or consortia of companies who are leaders in the industry being served. In the late 1990s third-party e-markets were the focus of a lot of new dot-com activity. Berlecon Research identified 2,500 independent exchanges on the Internet by 2000. Due to competition and the downturn in the economy, by late 2001 fewer than 150 of these exchanges were still in operation (Wichmann, 2002).

Covisint (http://www.covisint.com) is an example of a global marketplace sponsored by a consortium of industry leaders rather than an independent third party. In this case the industry leaders are DailmerChrysler, Ford, and General Motors among others. Covisint was jointly funded in 2000 to provide a global solution for the automotive industry with the goal to "improve the effectiveness of mission critical processes such as collaborative product development, procurement and supply chain management . . . through implementation and use of a secure and comprehensive online marketplace" (Covisint, 2002). Covisint's mission is to "connect the automotive industry in a virtual environment to bring speed to decision making, eliminate waste and reduce costs while supporting common business processes between manufactures and their supply chain."

Covisint provides three products: Covisint Auctions, Covisint AQP (Advanced Quality Planner), and Covisint Catalog. Covisint Auctions are online bidding events that provide rapid, real-time, Web-based negotiations in a secure environment for the purpose of supporting the sourcing of parts and components. Covisint AQP is an Internet-enabled application that provides an environment for collaboration, reporting, routing, and visibility of information important to developing high-quality standards for components being designed for vehicle production. Lastly, Covisint Catalogs are electronic purchasing environments for indirect and Maintenance, Repair, and Operations (MRO) material. This application allows users to shop online and provides a system to automate approvals and the creation of necessary supporting documentation such as purchase orders.

The total value of transactions managed by Covisint in 2001 was more than $129 billion. During the year, General Motors used Covisint to buy $96 billion worth of raw materials and parts for future vehicle models. In a single four-day period in May, DaimlerChrysler used the exchange to procure $3 billion worth of parts—the single largest online bidding event to date. Ford reported that it used the exchange to save $70 million, which is more than its initial investment in the exchange and expects to save approximately $350 million in 2002 (Konicki, 2001). Recent research has questioned the validity of the savings claims made by exchanges in general as they reflect the maximum theoretical savings that could be achieved at the close of the auction while the amount of actual savings is likely to be less (Emiliani & Stec, 2002).

METHODS FOR IMPLEMENTING B2B

Integrating a business process in one company with a business process in another requires that both companies provide the necessary technology that allows the information systems to be integrated over a computer network such as the Internet. Depending on the type of activity, this could include interfacing with other companies' supply-chain management, procurement, or sourcing systems, or involve something less complex such as simply establishing an e-mail system or sharing electronic files on tape or other digital media. Newer businesses are more likely to have modern information systems specifically designed for integration and collaboration; older companies are more likely to have legacy systems, which were never designed to be used for data sharing and pose significant challenges for B2B. A variety of methods that allow companies to overcome issues with legacy and incompatible business systems and allow them to effectively share business information have been developed.

Point-to-Point Transfer Methods

The point-to-point transfer of business information using computer data files dates back to the earliest B2B integration efforts. By the 1950s companies had already started to use computers to manage business transactions. Business information moving between companies at this time used paper forms, which had to be re-entered into the recipient computer systems using manual methods: a process that was unreliable, expensive, and redundant (Schneider, 2002). By the 1960s companies had started to use magnetic tape and punch cards as a medium to record business information for transfer. The encoded tapes or card decks would be transferred to a recipient computer system and automatically processed using card or tape readers. The advantages of these automated methods were in the removal of the expensive and error-prone process of redigitizing the business information as the data could be entered directly from the transfer medium.

Tape and card decks were replaced by automated file-transfer methods such as Tymnet, UUCP, Kermit, and later by the ubiquitous file transfer protocol (FTP) for use on the Internet. Using these applications, one computer system exports business data from an application to a data file; the file is then transferred to the recipient computer system over a phone line, computer network, or the Internet. At the recipient end the file is imported into the application using specialized software. FTP is part of a suite of protocols specifically designed for the Internet and is implemented today by most computer systems, allowing for a seamless transfer of binary or text data from one computer system to another. This “built-in” ease of use has largely replaced other file transfer applications for the business-to-business transfer of information.

The point-to-point business integration model using file transfer methods is simple to design and implement and works well in situations where there is little change in the applications sharing information and a high-degree of collaboration between the business entities. Yee & Apte (2001) identify several disadvantages of this approach. With respect to the format of the data being shared: the sending and receiving applications must agree on a fixed file format, export and import logic is often restricted by the applications, and lastly, the overall system is brittle and can fail if either system changes the data format. From a data management perspective, data transfer is conducted in batch mode rather than real time, which introduces latency into the system; there are no methods for recovery or guaranteed delivery of the data. Sharing data with multiple business partners becomes difficult to manage as the impact of these disadvantages becomes multiplied with each additional business partner.

Database Integration Methods

Point-to-point file transfer methods are perhaps the simplest integration approaches where the only factors in common between the collaborating systems are a prenegotiated file format and data content and a common communication method. Accessing data directly in the databases is another data integration method that can be relatively straightforward to implement given the correct set of preconditions.

Database software such as Oracle’s Enterprise Database, Microsoft’s SQL Server, and IBM’s DB2 software can be thought of as layered applications. At the lowest layer is the data itself and a set of procedures designed to efficiently manage the organization of the data on the host computer system. At the top level is a user interface, which allows users to connect to the database, manage the data, and format reports. Between the data and the user interface layers exists a suite of applications that implement the data management and query capabilities of the database. This application layer typically uses a specialized computer language called structured query language (SQL) to manage all aspects of a database and to interface between the users and the data itself. Database connectivity middleware such as JDBC, ODBC, and ADO allow direct programmatic access to the application layer of remote databases using a variety of programming languages and environments and replaces the user-interface layer of the database by generating SQL commands directly from the programming environment. This direct access allows systems designers to create points of direct integration between an application and a database over a telecommunications system such as the Internet. These database connectivity technologies also tend to shield the collaborating applications from the specifics of the data storage, overcoming differences in storage formats between disparate databases and computer systems. For example, a business transaction could be extracted by a remote application from an Oracle database instance on a Windows Server and inserted into an IBM DB2 database instance on an OS/400 server without difficulty, only relying on the understanding of the organization of the recipient database and permission to access the requisite database resources.

Newer business systems tend to decouple the application logic from the storage of the business data and store the data in a commercial database. This design makes the process of integration relatively simple as the database can be accessed directly. Older business systems, however, tend to tightly couple the business logic to the data and are less “open” in systems terms. These types of systems may use proprietary data stores and/or data storage formats, which makes integration at the database level problematic. Linthicum (2001) suggests that integration with these older “closed” systems is best managed at the application level as it is often impossible to deal with the database without dealing with the application logic as well.

A recent trend among database vendors is to facilitate interapplication data integration by allowing database software to send and receive data using XML syntax. XML data can be processed by the databases and either stored and queried in native (XML) form or converted to more traditional database storage types. This approach further serves to open up information systems for integration and often provides a way in which legacy systems with closed database platforms can integrate with newer systems.

Many companies use multiple databases for storing and maintaining their business data. Each database could be associated with a single application or be a point of integration for two or more applications in use within the company. Integrating data from multiple data sources provides a complication for normal direct database integration approaches as these typically require a single data source to connect with. To overcome the issue of integrating a business system with multiple target data sources database gateways can be used. Database gateways are middleware technologies that provide a query facility, often using SQL, against multiple target data sources. The middleware acts as a proxy and accepts requests for data from client systems, which it then translates into a form that can be executed against one or more connected databases. The database gateway middleware merges the underlying databases to form a "composite" or "virtual" database, which is a conjoining of all or selected parts of the underlying (managed) database schemas. IBM's Distributed Relational Database Architecture (DRDA) is an example of a database gateway built into IBM's DB2 enterprise database systems to facilitate interoperation of multiple databases within heterogeneous computing environments.

While proficient at providing read-only access to business data for remote applications, Yee & Apte (2001) maintain that database gateways systems have limitations for e-commerce systems. In particular, database gateways are inefficient when integrating multiple disparate systems as queries against the virtual database must be recast to query the underlying data sources and the results merged by the middleware to form a set of results that can be returned to the client. Further, Yee & Apte argue that the database approach to integration bypasses the business rules implemented in the application code and may result in redundant business logic, which must be both developed and maintained at some cost.

API Integration

As an alternative to integration through direct access to the data, applications can often share information via a set of interfaces "built in" to the application and designed to be accessed using other programs. These application programming interfaces (APIs), as they are called, allow external applications (clients) to share data and business logic with a host application (server) often via the Internet or other connection media. Depending on the type of technology used, the applications, and the form of the APIs, the two independent programs can share simple computer data composed of text strings and numbers and even complex structured data objects that represent an atomic information component such as a customer or purchase order. Some APIs allow the client system to execute business logic functions on the server such as removing an element of data or performing some action.

Integration via APIs is often more difficult in a heterogeneous computing environment as differences in how systems can be accessed and the representation of data components such as strings, numbers, dates, and complex higher-order objects can vary enough to render data generated on one system intelligible on another. These well-known system compatibility issues often require middleware applications, which can broker between the data representations on the different systems. Remote procedure calls (RPCs) are a middleware technology that provides a framework for translating system-specific data to and from a common data format, which can then be transferred without loss of representation between client and server systems over a computer network. RPC frameworks, initially developed to network UNIX applications together, are available on most computer systems used by businesses today and rely on a common computer language called interface definition language (IDL), which all data must be translated to before it can be sent over the network. Other API technologies are in common use. Microsoft (http://www.microsoft.com) has extended its Component Object Model (COM) to allow computers to share data and methods between systems running different versions of Windows software over a network. This Distributed COM, also known as COM+, is available on all second-generation Microsoft Server products. The Common Object Request Broker Architecture (CORBA) is a system similar to Microsoft's DCOM that was developed as an open specification by the Object Management Group (http://www.omg.org), a consortium consisting of over 800 independent software developers. The initial CORBA specification was developed in 1991 and, although it has been overshadowed in recent years by Java and XML, is still in use especially in the banking industry, who has adopted it as a standard method of integration.

Message-oriented middleware (MOM) was developed to address some of the issues associated with tightly coupled solutions such as RPC and COM. MOM applications transfer data between applications in the form of messages: an application-defined unit of data that could be in binary format or text-based. Most message-driven applications use an asynchronous processing model rather than the synchronous model used by most tightly coupled systems. In an asynchronous message-driven system one application sends messages to another application. Unlike RPCs, which deliver the data immediately, the messages sent from one application to another using MOM are typically placed into a specialized piece of software that allows the messages to be stored and queued. When the recipient application is ready to process messages it accesses the queue and extracts the messages from the queue that are for the application. This model allows the sending system to send messages to possibly several systems at the same time without having to wait for the recipient applications to process the data. This "fire-and-forget" model allows all systems to work independently and at different speeds. IBM's WebSphere MQ software—formally MQSeries—is an example of a widely deployed MOM system.

Java is a computer language developed by Sun Microsystems (http://java.sun.com) in the 1990s specifically for use on the Internet. Unlike most other computer languages Java uses a data format that is common to all Java applications, is independent of the hardware/software environment, and negates the need for an intermediate language such as IDL to represent data that will be transmitted between applications over a network. This "one-size-fits-all" approach greatly decreases the cost of developing and maintaining networked applications although Java critics maintain that Java is slow compared to other programming languages due to its device-independent run-time architecture. Since its inception the Java language has been rapidly growing to accommodate the evolving needs of the business community. Versions of the Java language are now available specifically to meet the needs of enterprise applications and mobile users. Java 2 Enterprise Edition (J2EE) provides a framework for developing robust distributed applications using specialized software platforms called application servers. J2EE application servers such as BEA Systems WebLogic (http://www.bea.com/framework.jsp?CNT = index.htm&FP = /content/products/platform), IBM's WebSphere (http://www-3.ibm.com/software/info1/websphere/index.jsp), and Oracle's 9iAS (http://www.oracle.com/ip/deploy/ias/) provide frameworks for efficiently accessing databases, managing transactions, creating message-driven applications, and developing Web-based interfaces. Java 2 Mobile Edition (J2ME) is a version of the Java language specifically designed for mobile devices such as Personal Digital Assistants (PDAs) or in-vehicle systems.

The recent adoption and proliferation of XML and XML-oriented middleware has provided an alternative means for sharing data between systems and is the basis of the new generation of interapplication methods called Web services. Web services are similar to RPCs in that middleware exists on both the client and the server that communicate and act as proxies for the client and server applications. Web services use a common XML document format based on the simple object access protocol (SOAP) specification developed jointly by Microsoft, IBM, and others to share data and can utilize commonly available Internet network connections and protocols such as HTTP and TCP/IP to communicate. Web services further benefit system developers because they can implement dynamic discovery mechanisms and centralized service registries that allow client applications to locate services and "discover" the interfaces and data formats available from the services and automatically integrate with them, thereby reducing the amount of time and the complexity required to build integration components.

Process-Oriented Integration

Process-oriented integration focuses on the logical sequencing of events and the processing of information as it moves within and between business organizations and relies on a business process rather than a technological foundation. The goal of process-oriented integration is for trading partners to share relevant business data with the intention of increasing competitive advantage through cooperation. This approach tends to be more strategic than tactical as the results are often hard to measure in terms of traditional investments as they involve developing trust relationships with suppliers and sharing private and often confidential information to realize more intangible benefits such as better products, increased customer satisfaction, and better supply-chain operation. Sharing production forecasts and schedules with suppliers, for example, allows business partners to better plan their own production activities, which in turn can lead to lower costs overall as the guesswork involved in anticipating demand can be removed and the likelihood of stock-outs diminished. Changes in production schedules can similarly be communicated, allowing suppliers to automatically adjust their production schedules to match, thereby reducing waste and uncertainty.

In its simplest form process-oriented integration might simply be a group of companies agreeing on a common suite of products to use for their internal systems such as SAP/R3 (http://www.sap.com) or Oracle's e-business applications suite (http://www.oracle.com/applications) and then deciding which processes they are willing to externalize as integration points. Establishing a common technology platform for the business systems also establishes a framework for sharing information as the data moving between businesses are guaranteed to be compatible and interchangeable as most enterprise-scale systems have built-in APIs or messaging systems to facilitate the sharing of data within and across organizations.

Evaluating and Selecting Integration Approaches

The wide variety of methods available for B2B system integration provides companies in the planning phase of a B2B initiative a number of alternatives they can choose among. Key design characteristics that must be evaluated for new systems are open vs closed implementations, integration at a data level using databases and/or middleware APIs or at a systemic level by adopting common business processes and systems, whether to use the public Internet as the data transport medium or to invest in private or semiprivate networks, and lastly, the ease of integration with existing and planned internal systems.

One of the key decisions is to determine the level of effort necessary for trading partners to couple and decouple from the trading environment. Systems such as SAP/R3, which use proprietary technology and require specialized hardware and connectivity environments, can be expensive and difficult to implement. Hershey Foods, for example, spent three years and $115 million implementing a software system from SAP. The new system was designed to replace a number of older systems and tie into supply-chain-management software from Manugistics and customer relationship management products from Siebel. Glitches in the ordering and shipping systems, however, resulted in a 12.4% drop in sales during the company's busiest quarter (Osterland, 2000).

ERP II is a new generation of Internet-centric enterprise resource planning (ERP) software designed specifically to address the types of implementation issues associated with sharing information across a supply chain experienced by Hershey Foods. For companies who have already invested in ERP systems, ERP II will allow them to leverage their investments and move toward a collaborative planning system by upgrading their existing ERP systems over time (Bond et al., 2002). Companies without an existing investment in ERP will be able to adopt a system that provides both internal and external business process integration.

In contrast to ERP II, light-weight technologies like XML/SOAP can provide points of integration between business systems and facilitate relatively low costs of entry into a trading consortium. A possible disadvantage of this relatively low cost of entry and ease-of-implementation is that trading partners could move to a competitor system with relative ease, thereby undermining the trust relationship between companies.

Other key decisions will center on the robustness, security, and scalability of the technology being selected. Last but not least, companies should examine the record of systems vendors and critically examine the vendors' record with respect to implementation success, after-sale support, and case studies proving the advertised ROI.

B2B E-COMMERCE CHALLENGES

There are several challenges to companies planning an e-B2B strategy, including managing and valuing e-B2B projects and getting up to speed on the regulatory environment surrounding e-B2B and the technical challenges associated with selecting and implementing e-B2B technologies.

Management Challenges

One significant management challenge associated with developing and sustaining a viable e-B2B initiative is measuring the true tangible and intangible benefits of the investment. Intangible benefits are hard to isolate and quantify and can accordingly effect how an investment is perceived by the company and shareholders. Other challenges for management include developing and maintaining an information technology strategy in an environment that is rapidly changing and evolving and managing the trust and expectations of business partners.

Measuring Intangible Benefits

Businesses choose to develop an e-B2B strategy for a variety of reasons, such as increasing process efficiency, reducing costs, and integrating new suppliers and customers. Tangible benefits such as increased sales, decreased production time, and reduced waste can be estimated fairly accurately. However, the intangible benefits of e-B2B are hard to evaluate using purely economic measures. For example, intangible benefits such as increased customer satisfaction and stronger relationships with business partners are two benefits that are highly sought but difficult to measure. The potential benefits from increased global visibility through an electronic presence on the Internet might include better hiring opportunities or more favorable investment opportunities. Again, such benefits are extremely difficult to quantify.

Managing Trust Relationships

Coupling business processes through technological frameworks requires that participants trust each other with valuable and often confidential business information such as new product specifications, purchasing patterns, and production forecasts. As businesses increasingly move toward a pattern of sharing information in real time a bond of trust must be established and proactively managed to ensure continued and mutual benefits for all companies involved. This trust relationship involves not only guaranteeing the security and confidentiality of the data being shared but also guaranteeing the accessibility and reliability of the systems being integrated. Managing trust relationships involves determining and maintaining predefined levels of system performance such as the speed and volume of transactions that can be processed, the stability of the information systems, and how systems should respond to erroneous or unprocessable data. Also important are agreements to establish responsibility and availability of the systems for routine maintenance and upgrade cycles.

Managing Information Technology Infrastructure

In an e-business, the information technology infrastructure of a company is the foundation for business success. The information technology infrastructure must be carefully managed to support the business goals of the company and should be perceived internally as a strategic asset. The Hurwitz Report (Hurwitz Group, 2001) identifies several ways in which the management and perception of the role of a company's IT organization is critical to success. In traditional companies, the IT group is responsible for managing internal processes and is used to support or maintain business functions. While IT groups are often treated as overhead and run as cost centers in traditional businesses, the IT organization in successful e-businesses is viewed as a revenue generator and treated as a competitive asset allowing innovation within the organization to drive competitive advantage and help reduce costs.

The reliability of the infrastructure is also paramount and must be carefully managed. As companies become more reliant on online processes, the potential consequences to a company and its business partners of even a temporary loss of service can be devastating. For example, a survey of companies in 2001 showed that for 46% of the companies surveyed, system downtime would cost them up to $50k per hour, 28% of the respondents would incur a loss of up to $250k per hour, and 26% would loose up to or over $1m per hour (Contingency Planning Research, 2001).

Managing Expectations

Simply developing the e-business technology is just the start of the electronic collaboration process. A key challenge to successfully implementing an e-business strategy is managing the expectations of the e-business and effectively communicating the benefits to business partners. At the request of the Boards of Grocery Manufacturers of America and Food Marketing Institute Trading Partners Alliance (TPA), A. T. Kerney developed an action plan to accelerate the degree of cooperation between TPA members after over $1 billion had already been invested in a variety of exchanges and electronic collaboration platforms within the industry. Central to the recommendations made by A. T. Kerney was the need for better communication among the partners to address common concerns over data synchronization, education about the benefits of collaboration and implementation best practices, and regular feedback through surveys and progress tracking initiatives. Also identified was the need for individual companies to proactively encourage trading partners to join through training and sharing of best-practices (Kerney, 2002).

Monitoring and Regulation Challenges

There are many challenges associated with monitoring and regulating online businesses. These challenges are exacerbated by the increasing internationalization of business. Issues with taxation, security, and privacy are more difficult to manage when applied in a global environment where the laws and ethics governing business are often conflicting rather than complementary. Issues also exist with standardizing the accounting mechanisms for digital businesses and processes as traditional accounting principles must adapt to the new business environment. Lastly, due to the global nature of the Internet, issues with managing the security of the digital systems and prosecuting those who attempt to disrupt the flow of digital information provides significant challenges for international legal organizations.

Internet Business

As a global phenomenon, e-B2B poses a complicated series of issues for those organizations charged with monitoring and regulating international business. The Internet compresses the natural geographic separation of businesses and the related movement of products and money between these businesses and allows business operations to span political borders literally at the speed of light. To date, the Internet is largely uncontrolled and business on the Internet follows suit. It is the responsibility of individual countries to decide how to regulate, how much to regulate, and who should regulate the Internet and Internet-related business operations within their boundaries, as well as the degree to which they should accommodate the rules and regulations set up by other countries.

Within any country there are opposing forces at work. Governments are trying to regulate the Internet in a way that stimulates their economies and encourages use. At the same time governments must protect the rights of citizens and existing businesses. It is not surprising that few Internet regulatory laws have been passed even though e-B2B has been in place for many years. To further complicate the regulatory issue, the ubiquity of the Internet has resulted in the rapid expansion of businesses interoperating across international borders. Traditionally only large, well-financed companies could perform international commerce; now there are no limits to how small a multinational company can be.

Confidentiality and Privacy

As businesses and industries in general move toward higher degrees of collaboration at a digital level, issues about the security of the digital information being collected, stored, and transmitted as part of normal business operations becomes more important. Many of the regulatory laws that do exist pertain to protection of personal, confidential, or legally sensitive business data. For example, in the United States personal finances and healthcare records now must be protected from accidental or malicious exposure. The Health Insurance Portability and Accountability Act (HIPAA, 1996) is an attempt to regulate the movement, storage, and access to healthcare-related personal information through enforcement of privacy standards. HIPAA is a direct regulatory response within the United States to several well-publicized breaches of doctor–patient trust including the errant transfer of prescription records on a computer disk sold to an individual in Nevada, medical records posted on the Internet in Michigan, and digital media containing medical records stolen from a hospital in Florida. These are just a few of the better-publicized cases. The HIPAA regulation is scheduled to go into effect in 2003.

Accounting for Digital Transactions and Digital Assets

The Federal Accounting Standards Board (FASB, 2001) identifies several challenges to the accounting industry associated with the transition from a traditional paper-driven economy to one where business is transacted using digital documents. From an accounting perspective, there is no standard measure for determining the short-term or long-term value of technology, knowledge capital, and intellectual property associated with implementing and managing an e-business infrastructure. Yet companies are investing huge amounts of capital in these activities that must be accounted for fairly. Valuing these intangible assets would require that the accounting profession extend its province to nonfinancial areas of business operations and generate standard frameworks and metrics for reporting and tracking nonfinancial information.

Computer Crimes

There are many types of criminal activity associated with the Internet, ranging from petty acts of vandalism and copyright violations to organized systematic attacks with malicious intent to cause damage, financial loss, or in extreme cases, wholesale disruption of critical infrastructure. Each week it seems that there are reports on new attacks, break-ins, viruses, or stolen data made public. Different countries, however, treat different types of computer attacks in different ways with little consensus as to the degree of criminality and severity of punishment. Turban et al. (2001) provide a comparison of computer crime legislation in several countries and note that an activity such as attempting to hack (to gain access to an unauthorized system), while legal in the United States, is illegal in the United Kingdom. However, successful hacking, which causes loss, is criminal in both countries but punished far more severely in the United States, carrying a maximum penalty of 20 years.

Security breaches and malicious attacks such as vandalizing Web sites and disabling computer servers using viruses, denial-of-service (DoS) attacks, and buffer-overflow techniques can prove costly to e-business by making the systems unavailable for normal operation, often resulting in incomplete transactions or lost business while the systems are incapacitated. A recent survey based on 503 computer security practitioners in the United States reported that 90% of respondents detected security breaches in their systems within the 12-month reporting period. Eighty percent of the respondents attributed financial loss to these security breaches, amounting to approximately $500 million from the 44% who were able to quantify their loss (Computer Security Institute, 2000).

The global nature of the Internet makes regulating this so-called cybercrime a significant issue as attacks against systems can be launched from countries with less-stringent regulations and lower rates of enforcement. The United States has recently caused international concern by successfully detaining and convicting foreign nationals accused of crimes against systems in the United States even thought the attacks were launched from non-U.S.-based systems (U.S. Department of Justice, 2002). However, not all attacks are malicious and are often perpetrated to expose security holes in business systems as a means to draw public attention to issues of vulnerability with the intent of eliciting a general hardening of the systems by the systems developers (Hulme, 2002).

Technological Challenges

Some of the most extreme information technology requirements found in the commercial business world are associated with designing and implementing large-scale e-business systems. Issues with the design and deployment of such systems include localizing the applications for international users, managing the quality of service, and protecting access to the data and applications.

Localization and Globalization

Like other forms of e-commerce, e-B2B is increasingly multinational and differences in language and culture can affect the usability of systems. One technological challenge is the process of designing and developing software that functions in multiple cultures/locales, otherwise known as globalization. Localization is the process of adapting a globalized application to a particular culture/locale.

A culture/locale is a set of rules and a set of data specific to a given language and geographic area. These rules and data include information on character classification, date and time formatting, numeric, currency, weight, and measure conventions, and sorting rules. Globalizing an application involves identifying the cultures/locales that must be supported by the application, designing features that support those cultures/locales, and developing the application so that it functions equally well in any of the supported cultures/locales.

If business systems are to share information across national and cultural borders, the information they are sharing must be both syntactically correct and unambiguous for all the systems involved. Of particular importance to cross-cultural systems are the design and development of data storage and data processing systems that can accommodate the differences in data and translate the data between supported locales. Consider three companies operating in the United States, U.K., and Japan respectively, and sharing data about a shipment date that will occur on the date specified by the string "07/08/02." In the United States this date string is interpreted as July 8, 2002. In the U.K. the date and month field are transposed, resulting in a local/culture-specific interpretation of August 7, 2002. In Japan the year and month field are transposed resulting in an interpreted shipment date of August 2, 2007. This is a simple example but the cost associated with developing new or re-engineering existing systems to translate culture/locale-specific values such as simple date strings across a variety of culture/locales is high. There are similar issues associated with sharing data that contains time information, addresses, telephone numbers, and currency.

Scalability, Reliability, and Quality of Service

The general planning considerations for engineering a service such as a B2B marketplace are to provide sufficient system functionality, capacity, and availability to meet the planned demand for use of the systems, which translates loosely into the number of transactions that can be processed in a given time period. The difficulty lies in estimating and planning the processing requirements for a system and engineering a system that can respond to periodic increases in use. In particular, system designers need to understand the demand on the system during "normal" operations and how the demand on the system might vary over a period such as a single day, a week, a month, or the course of a year. Further, system designers need to plan for the effects of "special events," which might cause a short-term increase in the use of the system. Lastly, the system designers need to understand how the B2B interactions translate to workload against the internal systems such as an ERP and plan for matching the capacity of the internal systems with that of the external-facing systems being designed so that the internal systems do not become the weak link in the processing chain.

From a trading-partner perspective, the primary concerns will pertain to the quality of service (QOS) provisions for the system they are going to integrate against. The QOS concerns can be broken down into four key areas: system reliability, system security, system capacity, and system scalability. E-business companies need to set design goals for these key QOS requirements and plan the level of investment and predicted ROI accordingly.

Protection of Business Data and Functions

One result of growth in e-business activity is the associated increase in the transmission and storage of digital information and the corresponding increase in reliance on information systems to support business activities. This poses two major problems to information technology management: how to maintain the integrity and confidentiality of business information and, secondly, how to protect the information systems themselves from security breaches, malicious attacks, or other external factors that can cause them to fail.

Potentially sensitive information shared between businesses is at risk not only during the transmission of the information from one system to another but also as it is stored on file servers and in databases accessible over the computer network infrastructure. The general trend toward open systems poses transmission security issues because open systems rely on text-based data formats such as EDI and XML. If intercepted, these documents can be read and understood by a variety of software publicly available on the Internet. Standards for encrypting XML documents are being developed but have yet to make the mainstream. Encryption technologies for securing data moving over the Internet between business partners, such as VPNs, have been available for a decade or more but rely on coordinating privacy schemes among businesses, sharing encryption keys, and above all developing effective implementation policies that must be constantly revised and tuned to adapt to changing events. Even using encryption systems is often not enough as systems using 40- or 56-bit keys can now be broken using brute-force methods in a few minutes using a personal computer. On the other hand, strong encryption systems that protect data from these brute-force attacks can often not be shared with partners outside of the United States.

Maintaining the integrity and confidentiality of business data and computer systems is important. An often overlooked part of the e-business process, however, is the protection of access to the business systems themselves. The tragic events of September 11, 2001, in New York City have served to underline the fragility of electronic business systems and have provided new priorities for e-business companies. Rather than just simply protecting business data stored in corporate databases by backing it up on tape, companies are now considering holistic approaches to protecting business operations, including protecting access to critical applications such as e-mail as well as protecting lines of communication to those businesses upon which they are dependent (Garvy & McGee, 2002).

B2B E-COMMERCE IN PERSPECTIVE

Whether or not the actual global value of e-B2B activity meets the predictions for 2005 mentioned in the Introduction, the current rates of adoption and continuing pervasiveness of e-B2B activity across all types of business and all types of industry are indicators that e-B2B is going to remain a major driving force within the global economy. At the same time, e-B2B should be regarded as a nascent activity that is rapidly emerging and consequently exhibiting growing pains. Based on a survey of 134 companies around the world, the Cutter Consortium provides some interesting insight into the e-B2B implementation experience (Cutter Consortium, 2000):

Asked to rank the obstacles to e-business, respondents chose “benefits not demonstrated” as the number-one obstacle, followed by financial cost and technological immaturity.

Success with electronic supply-chain management is mixed, with half of those using it enjoying success rates of 76–100% and about a third experiencing success rates of 0–10%.

Similarly, Deloitte & Touche examined data from 300 U.S.- and U.K.-based companies in a wide variety of industries and identified a highly conservative trend to e-business. Less than half of the companies examined expected their e-business strategy to involve transforming business processes while the majority expected to engage in simple Internet-based buying and selling. More telling is that only 28% of the companies had actually developed a formal e-business strategy, the majority being in some stage of investigation (Rich, 2001).

Undemonstrated returns on investment, unsuccessful implementations, and lack of knowledge about the technology, the risks, and the rewards associated with e-B2B all combine to make businesses cautious as they plan ahead for a connected future. The Deloitte & Touche study reports that approximately 50% of respondents indicated that the major barriers to e-B2B lie in the lack of skills and training, or existing business culture compared to approximately 20% of respondents who perceive technological or security issues as the major barrier. These are issues that can be overcome with time as the global corporate knowledge base grows, as more case studies illustrating successful and profitable implementations are available, and as the technology framework becomes more robust, secure, and prevalent.

The success stories available are highly compelling: General Motors, Federal Express, and Cisco are examples that clearly demonstrate the returns possible when e-business is implemented successfully. The bottom line that drives most organizations to evaluate and adopt e-B2B is that performing business electronically is both cheaper to execute and more efficient in terms of time than traditional means. An electronic transaction is also more accurate, which often draws trading partners into supply-chain integration efforts. Indeed, recent evidence indicates that businesses are still investing in e-B2B infrastructure, particularly for electronic supply-chain integration including e-procurement as well as for e-markets. Berlecon Research estimate that the number of B2B exchanges worldwide will grow 10-fold over the next two years particularly in Europe where there are several underserved industries that lack online B2B marketplaces (Wichmann, 2002).

Clearly, implementing an e-business strategy is highly technical and involves many facets of information technology that are new to most companies and indeed new to the information technology industry as well. In essence, everyone is learning how to perform e-business and the technical solutions are adapting to meet the evolving requirements of e-B2B. However, as one author puts it, “Business comes before the e” (Horne, 2002), reinforcing the idea that e-business is not a means to an end but should be regarded as an extension of a sound and well-constructed business plan.

GLOSSARY

ActiveX Data Objects (ADO) Middleware developed by Microsoft for accessing local and remote databases from computers running a Windows-based operating system.

Application programming interface (API) Components of an application that allow other applications to connect to and interact with the data and services of the application; usually published by the application developer as a formal library of functions.

Dot-com Companies that emerged during the Internet boom of the late 1990s with a focus on building applications for or selling products on the World Wide Web; usually funded by large amounts of venture and private equity capital.

E-business All types of business activity performed using electronic means; has a wider context than e-commerce and includes business exchanges that involve interorganizational support, knowledge sharing, and collaboration at all levels.

E-commerce A form of e-business that results in a business transaction being performed using electronic means such as a buy or sell event.

Electronic data interchange (EDI) One of the first forms of e-business used to pass electronic documents between computer systems often over private telecommunication lines or value-added networks; still used by many businesses today.

Electronic funds transfer (EFT) An early form of e-commerce used to transfer money electronically between banks.

Enterprise resource planning (ERP) A collection of applications designed to manage all aspects of a business. ERP systems are designed to integrate sales, manufacturing, human resources, logistics, accounting, and other enterprise functions within an organization.

Intermediary A company that adds value or assists other companies in performing supply-chain activities such as connecting suppliers with buyers.

Java Database Connectivity (JDBC) Middleware used for accessing remote databases from an application written using the Java programming language.

Message oriented middleware (MOM) Middleware used to integrate applications using a system of messages and message queues; allows systems to share data using an asynchronous processing model.

Middleware Software that enables the access and transport of business data between different information systems often over a computer network.

Open Database Connectivity (ODBC) Middleware specification originally designed by Microsoft to access databases from Windows platforms using a standard API. ODBC has since been ported to UNIX and Linux platforms.

Protocol A standard for the format and content of data passed between computers over a computer network; often maintained by independent organizations such as the World Wide Web Consortium.

Remote procedure call (RPC) Middleware technology originally developed for UNIX systems for sharing data and methods between applications over a network.

Structured query language (SQL) A computer language developed by IBM in the 1970s for manipulating data stored in relational database systems; became the standard language of databases in the 1980s.

Supply chain The end-to-end movement of goods and services from one company to another during a manufacturing process.

Transmission control protocol/Internet protocol (TCP/IP) Communication protocols originally developed as part of ARPANET that today form the basic communication protocols of the Internet.

Transaction A record of a business exchange such as a sell or a buy event.

Tymnet An early value-added network developed by Tymshare Inc. and used by companies for transferring computer files between computer systems; the largest commercial computer network in the United States and was later sold to MCI.

UNIX-to-UNIX-Copy (UUCP) A utility and protocol available on UNIX systems that allows two computers to share files over a serial connection or over a telephone network using modems.

Value-added network (VAN) A form of computer network connection often between two companies to perform e-business that is managed by a third party; initially used to transfer EDI documents; modern ones can operate over the Internet.

Virtual private network (VPN) A form of network connection between two sites over the public Internet that uses encrypted data transmission to provide a private exchange of data.

CROSS REFERENCES

See *Business-to-Business (B2B) Internet Business Models; Business-to-Consumer (B2C) Internet Business Models; Click-and-Brick Electronic Commerce; Collaborative Commerce (C-commerce); Consumer-Oriented Electronic Commerce; Electronic Commerce and Electronic Business; Electronic Data Interchange (EDI); Electronic Payment; E-marketplaces; Internet Literacy; Internet Navigation (Basics, Services, and Portals).*

REFERENCES

Aberdeen Group (2001a). *FedEx taps e-procurement to keep operations soaring, cost grounded*. Retrieved November 16, 2002, from http://www.ariba.com/request_info/request_information.cfm?form=white_paper

Aberdeen Group (2001b). E-procurement: Finally ready for prime time. Retrieved November 16, 2002, from http://www.aberdeen.com/ab_company/hottopics/eprocure/default.htm

Bond, B., Genovese, Y., Miklovic, D., Wood, N., Zrimsek, B., & Rayner, N. (2002). ERP is dead—Long live ERP II. Retrieved January 18, 2003, from http://www.gartner.com/DisplayDocument?id=314701

Colkin, E. (2002, September 16). Hastening settlements reduces trading risk. *Information Week,* 24.

Computer Security Institute (2000). *CSI/FBI computer crime and security survey.* Retrieved November 16, 2002, from http://www.gocsi.com/press/20020407.html

Contingency Planning Research (2001). *2001 Cost of downtime survey.* Retrieved January 18, 2003, from http://www.contingencyplanningresearch.com

Cutter Consortium (2000). *E-business: trends, strategies and technologies*. Retrieved November 16, 2002, from http://www.cutter.com/itreports/ebustrend.html

Davis, W. S., & Benamati, J. (2002). *E-commerce basics: Technology foundations and e-business applications*. Boston: Addison Wesley.

Emiliani, M. L., & Stec, D. J. (2002). Aerospace parts suppliers' reaction to online reverse auctions. Retrieved January 18, 2003, from http://www.theclbm.com/research.html

Financial Accounting Standards Board (FASB) (2001). *Business and financial reporting, challenges from the new economy* (Financial Accounting Series Special Report 219-A). Norwalk, CT: Financial Accounting Foundation.

Gartner Group (2001). *Worldwide business-to-business Internet commerce to reach $8.5 trillion in 2005*. Retrieved November 16, 2002, from http://www3.gartner.com/5_about/press_room/pr20010313a.html

Greenstein, M., O'Leary, D., Ray, A. W., & Vasarhelyi, M. (in press). *Information systems and business processes for accountants*. New York: McGraw Hill.

HIPAA (1996, August 21). *Health Insurance Portability and Accountability Act of 1996, Public Law 104-191*. Retrieved November 16, 2002, from http://aspe.hhs.gov/admnsimp/pl104191.htm

Horne, A. (2002). *A guide to B2B investigation*. Retrieved November 16, 2002, from http://www.communityb2b. com/news/article.cfm?oid=620910D2-91EF-4418-99B311DBB99F937B

Hulme, G. (2002, August 8). With friends like these. *Information Week*. Retrieved November 16, 2002, from http://www.informationweek.com/story/IWK20020705S0017

Hurwitz Group (2001). *E-business infrastructure management: The key to business success*. Framingham, MA: Hurwitz Group.

Garvy, M. J., & McGee, M. K. (2002, September 9). New priorities. *Information Week*, 36–40.

Kerney, A. T. (2002). *GMA-FMI trading partner alliance: Action plan to accelerate trading partner electronic collaboration*. Retrieved November 16, 2002, from http://www.gmabrands.com/publications/docs/ecollexec.pdf

Konicki, S. (2001, August 27). Great sites: Covisint *Information Week*. Retrieved November 16, 2002, from http://www.informationweek.com/story/IWK20010824S0026.

Linthicum, D. S. (2001). *B2B application integration: E-business enable your enterprise*. Boston: Addison-Wesley.

Microsoft (2000). *MS market—Intranet-based procurement*. Retrieved January 18, 2003, from http://www.microsoft.com/technet/treeview/default.asp?url=/technet/itsolutions/intranet/case/msmproc.asp

Norris, G., Hurley, J., Hartley, K., Dunleavy, J., & Balls, J. (2000). *E-business and ERP: Transforming the enterprise*. New York: Wiley.

Osterland, A. (2000, January 1). Blaming ERP. *CFO Magazine*. Retrieved November 16, 2002, from http://www.cfo.com/article/1,5309,1684,00.html

Rich, N. (2001). *e-Business: The organisational implications*. Retrieved November 16, 2002, from http://www.deloitte.com/dtt/cda/doc/content/Man_nrebriefing.pdf

Schneider, G. P. (2002). *Electronic commerce* (3rd ed.). Canada: Thomson Course Technology.

Slater, D. (2002, April 1). GM shifts gears. *CIO Magazine*. Retrieved November 16, 2002, from http://www.cio.com/archive/040102/matters.html

Small Business Administration (SBA) (2000). *Small business expansions in electronic commerce: A look at how small firms are helping shape the fastest growing segments of e-commerce*. Washington, DC: U.S. Small Business Administration Office of Advocacy.

Strategis (2002). *Electronic commerce in Canada*. Retrieved November 16, 2002, from http://ecom.ic.gc.ca/english/research/b2b/index.html

Turban, E., King, D., Lee, J., Warkentin, M., & Chung, H. M. (2002). *Electronic commerce: A managerial perspective* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

U.S. Department of Justice (2002). *Russian computer hacker sentenced to three years in prison*. Retrieved January 19, 2003, from http://www.cybercrime.gov/gorshkovSent.htm

Wichmann, T. (2002). *Business-to-business marketplaces in Germany—Status quo, opportunities and challenges*. Retrieved November 16, 2002, from http://www.wmrc.com/businessbriefing/pdf/euroifpmm2001/reference/48.pdf

Yee, A., & Apte, A. (2001). *Integrating your e-business enterprise*. Indianapolis, IN: Sams.

Business-to-Business (B2B) Internet Business Models

Dat-Dao Nguyen, *California State University, Northridge*

NATURE OF B2B E-COMMERCE

In general, B2B e-commerce can be classified according to the nature of the goods/services in transaction, the procurement policy, and the nature of the supply chain.

Businesses conduct B2B e-commerce to sell or buy goods and services for production and/or nonproduction. *Production materials,* or direct materials, go directly to the production of goods or services. Usually they are not shelf items that could be readily purchased at anytime in the marketplace. Their use is scheduled according to a production plan. They are purchased in large volume after negotiation and contracting with the sources to guarantee a continuous stream of input materials for the production process. *Nonproduction materials,* or indirect materials, are used in maintenance, repairs, and operations. They are also called MROs, containing low-value items. Although constituting 20% of purchase value, they amount to approximately 80% of an organization's purchased items.

Depending on the strategic nature of the materials in the production, a procurement policy may involve a long-term contract or an instant purchase/rush order. S*trategic sourcing* for a long-term contract results from negotiations between suppliers and buyers. *Spot buying* for an instant purchase/rush order concludes at a market price resulting from the matching of current supply and demand. Due to the strategic role of direct materials in the production, a manufacturer wishes to secure a consistent long-term transaction at an agreed upon price with its suppliers. Most organizations spend a great deal of time and effort for upstream procurement of direct materials, usually high-value items, and overlook low-value items, including MROs. Consequently, there are potential inefficiencies in the procurement process, such as delays in production due to insufficient MROs and/or overpayment for rush orders to acquire these MROs.

From a value chain and supply chain perspective, B2B e-commerce can take place in a vertical market or a horizontal market. A *vertical market* involves transactions between and among businesses in the same industry or industry segment. Usually this market deals with the production process or direct materials necessary for the production of goods and services of firms in the same industry. A *horizontal market* involves transactions related to services or products of various industries. These materials are diversified and related to the maintenance, repair, and operation of a specific firm. Most of these materials do not contribute directly to the production of goods and/or services offered by the firms.

Before the advent of B2B e-commerce, companies used the following tendering process. A department in a company submits a requisition for the goods/services it needs. The purchasing agent prepares a description of the project, giving its specification, quality standards, delivery date, and required payment method. Then the purchasing department announces the project and requests proposals via newspaper/trade magazine ads, direct mail, fax, or telephone. Interested vendors/suppliers may ask for and receive detailed information by mail. Then these suppliers prepare and submit proposals. Proposals are evaluated by several departments/agents at the buying company. Negotiation may take place and the contract is awarded to the supplier who offers the lowest price. In this process, communication mostly takes place via letter or fax/phone.

Similarly on the selling side, a supplier announces in newspapers or trade magazines the inventory to be disposed of and invites interested parties to inquire and bid. A sales force may be set up to identify and make direct contact with potential buyers. The interested parties may request additional information about the goods to be delivered by mail or fax. Then they decide to submit a sealed bid by mail. At the closing date, the bids are examined and the highest bid wins the auction.

This manual and paper-based process takes a long time and is prone to error. Electronic processes conducted via telecommunication systems are faster but require an investment in a dedicated private network. Recently, Web-based technologies have made business communication much less expensive and easier to administer. Transactions over the Internet also make it possible to reach a larger pool of business partners and to locate the best deal for the project.

There are many classification schemes for B2B e-commerce business models (Laudon & Traver, 2001; Turban et al., 2002; Pavlou & El Sawy, 2002). The classifications provide details on the context and constraints of the various business models so that an interested business may select an appropriate strategic choice to gain a competitive advantage.

In the following, B2B e-commerce business models are classified on the basis of business ownership and transaction methods.

B2B BUSINESS MODELS BY OWNERSHIP

Depending on who is controlling the marketplace and initiating the transactions, B2B e-commerce can be classified as company-centric or an exchange model. A *company-centric model,* representing a one-to-many business relationship, involves one business party initiating transactions and deals with many other parties interested in buying or selling its goods and services. In a *direct-selling* model, a company does all the selling to many buyers, whereas in a *direct-buying* model a company does all the buying from many suppliers.

In these models, the initiative company has complete control over the supportive information systems. However, a third party may serve as an *intermediary* to introduce buyers to sellers and vice versa, and to provide them with a platform and other added-value services for transaction. In many cases, buyers and suppliers having idle capacity in their Internet host sites for B2B e-commerce have served as intermediaries for other smaller businesses.

An *exchange or trading model,* representing a many-to-many business relationship, involves many buyers and many suppliers who meet simultaneously over the Internet to trade with one another. Usually there is a market maker who provides a platform for transactions, aggregates the buyers and sellers, and then provides the framework for the negotiation of prices and terms.

A variation of the exchange model is a *consortium trading exchange,* which constitutes a group of major companies that provide industry-wide services that support the buying and selling activities of its members. The activities can be vertical/horizontal purchasing/selling.

Direct Selling

This is a company-centric B2B model focusing on selling, in which a supplier displays goods and services in a catalog at its host site for disposal. The seller could be a manufacturer or a distributor selling to many wholesalers, retailers, and businesses.

In this model, a large selling company transacts over a Web-based, private-trading sales channel, usually over an extranet, to its business customers. A smaller business may use its own secured Web site. The company could use some transaction models, such as direct selling from electronic catalogs, requests for proposal (RFP), and selling via forward auctions, and/or one-to-one dealing under a long-term contract. The classification of these transaction methods in B2B e-commerce is discussed in detail in the next section.

In the B2B direct selling, the involved parties may benefit from speeding up the ordering cycle and reducing errors processing. They also benefit from reducing order processing costs, logistics costs, and paperwork, especially the reduction of buyers' search costs in finding sellers and competitive prices and the reduction of sellers' search costs in advertising to interested buyers.

Most major manufacturers have conducted B2B e-commerce with their business partners. For example, *Dell.com, Cisco.com, IBM.com, Intel.com,* and *Staples.com,* among others, have special secured sites for registered partners to provide them with information on products, pricing, and terms. At this site, business customers can browse the whole catalog, customize it, and create and save a shopping list/shopping cart for internal approval before placing orders. The sites have the tracking facility for customers to follow up on the status of their orders. These sites also have links to shipper's Web site (UPS, FedEx, Airborne Express, etc.) to help customers keep track of delivery.

In this model, depending on whether a manufacturer/distributor that hosts its own Web site and support provides complete transaction or not, the company may have to pay fees and commissions to intermediaries for hosting and value-added services. Usually, corporate buyers have free access to the e-marketplace after a free registration to the site.

Direct Buying

This is a company-centric B2B model focusing on buying, in which a company posts project specifications/requirements for goods and services in need and invites interested suppliers to bid on the project.

In this model, a buyer provides a directory of open requests for quotes (RFQs) accessible to a large group of suppliers on a secured site. The buying company doesn't have to prepare requests and specifications for each of these potential tenders. Suppliers could be notified automatically with an announcement of available RFQs, or even the RFQs sent directly from the buyer site. Independent suppliers can also use search-and-match agent software to find the tendering sites and automate the bidding process. Then suppliers can download the project information from the Web and submit electronic bids for projects. The reverse auction could be in real time or last until a predetermined closing date. Buyers evaluate the bids and negotiate electronically and then award a contract to the bidder that best meets their requirements. A large buying company can also aggregate suppliers' catalogs at its central site for the ease of access from its own branch offices. These affiliations will purchase from the most competitive supplier. In this case, the suppliers will be notified directly with the invitation for tender or purchase orders.

This model streamlines and automates the traditional manual processes of requisition, RFQ, invitation to tender, issue of purchase orders, receipt of goods, and payment. The model makes the procurement process simple and fast. In some cases, it increases productivity by authorizing purchases from the units/departments where the goods/services are needed and therefore bypassing some paperwork at the procurement departments. The model helps in reducing the administrative processing costs per order and lowering purchase prices through product standardization and consolidation of orders.

The model also contributes to improving supply chain management by providing information on suppliers and pricing. It helps to discover new suppliers and vendors who can provide goods and services at lower cost and on a reliable delivery schedule. It also minimizes purchases from noncontract vendors at higher prices for uncontrollable quality goods/services.

An example for this model is GE's *Trading Process Network* (TPN), where a company's sourcing department receives internal material requests and sends off RFQs to external suppliers. Currently GE opens this network to other business partners, at *gxs.com* site.

In this model, a buying company may set up its own Web site and may engage other services from intermediaries with licenses. Suppliers have to register at the host site and may have to pay an access fee.

Exchange/Trading Mall

The exchange model involves many suppliers and buyers meeting at the marketplace for transactions. The marketplace could be a dedicated site or a trading mall open to the public. Transactions in this marketplace involve spot buying as well as negotiation for a long-term buying/selling contract. In spot buying, a deal is concluded at a price based on supply and demand at any given time at the marketplace. In systematic sourcing, the exchange aggregates the buyers and sellers and provides them with a platform for the negotiation of prices and terms.

In the exchange, a company lists a bid to buy or an offer to sell goods/services. Other sellers and buyers in the exchange can view the bids and offers, although the identity of the tenderer or the bidder is kept anonymous. Buyers and sellers can interact in real time, as in a stock exchange, with their own bids and offers to reach an exact match between a buyer and a seller on price, quantity, quality, and delivery term. Third parties outside the exchange may provide supporting services, such as credit verification, quality assurance, insurance, and order fulfillment.

The exchange provides an open marketplace so that buyer and seller can conclude/negotiate the transaction at a competitive price resulting from the supply/demand mechanism. It has the characteristics and benefits of a competitive market in terms of classic economics. A buyer may benefit from lower costs due to a large volume of goods/services being transacted. A supplier may benefit from reaching a larger pool of new buyers than is possible when conducting business in a traditional market.

In this business model, some exchanges act purely as information portals by transferring the order/inquiry to the other party via hyperlinks so that the transactions will take place at the seller/buyer sites. Others aggregate suppliers and/or buyers for the convenience of the trading parties. In supplier aggregation, the exchange standardizes, indexes, and aggregates suppliers' catalogs and then makes them available to buyers at a centralized host site. Or requests for proposals (RFPs) from participant suppliers are aggregated and matched with demand from participant buyers. In buyer aggregation, RFQs of buyers, usually the small ones, are aggregated and linked to a pool of suppliers that are automatically notified of the existence of current RFQs. Then the trading parties can make bids.

Another type of exchange is called a *consortium trading exchange,* formed by a group of buyers or sellers. In *buying consortia,* a group of companies joins together to streamline the purchasing process and to pressure the suppliers to cut prices and provide quality, standardized goods/services in vertical as well horizontal supply chain transactions. An example of a buying consortium is *Covisint.com,* an automotive industry joint venture by GM Motors, Ford, DaimlerChrysler, Renault, Peugeot Citroen, and Nissan. In *selling consortia,* suppliers in the same industry deal with other downstream businesses to maintain reasonable prices and controllable production schedules for goods/services in vertical trading. An example of a selling consortium is the *Star Alliance,* an alliance of major domestic and international airlines, consisting of Air Canada, Lufthansa, SAS, United Airway, and others. These companies sell or exchange seats in their airplanes to one another to assure full booking for their fleets.

The use of exchange would especially benefit smaller businesses, which don't have large customer bases or supplier sources. Transactions via exchange and intermediary sites don't require additional resources for information technology infrastructure, staffing, and other related costs. If the exchange is controlled by an intermediary, this third party often assumes the responsibility for credit verification, payment, quality assurance, and prompt delivery of the goods.

There are intermediary exchanges, such as *eBay.com* and the Trading Information Exchange of GE Global eXchange Services (*gxs.com*), that provide an open marketplace for many suppliers/vendors and buyers. In other cases, manufacturers provide upstream and downstream partners with a service enabling them to do business with one another.

One prominent example is Boeing's secured site *MyBoeingFleet.com,* at which Boeing's airline customers can access the PART page to order maintenance parts directly from Boeing's suppliers. This service significantly streamlines time and labor in the procurement process for all business partners. One no longer needs to go through archives to look for blueprints, specifications, and sources of thousands parts of an aircraft for the requisition of a specific item.

On the revenue models of exchanges, if a major partner of the supply chain owns the site, access to an exchange marketplace could be free of charge. In other cases, participants pay an annual registration fee and/or a transaction fee that is either a fixed amount or a percentage

Table 1 B2B E-commerce Business Models

BY TRANSACTION METHODS	BY OWNERSHIP		
	Direct Selling	Direct Buying	Exchange
Electronic Catalogs	•		
Automated RFQs		•	
Digital loyal networks	•	•	
Metacatalogs		•	•
Order aggregation		•	•
Auction	•	•	•
Bartering			•

of the transaction volume. The participants may also pay for added-value services, such as credit verification, insurance, logistics, and collection, provided by the exchange. Some exchanges generate extra revenue from online advertisements on the site.

B2B BUSINESS MODELS BY TRANSACTION METHODS

B2B business models could be classified by the transaction methods a buying/selling company uses to conduct business with its partners in the e-marketplace. A company may use one or many transaction models suitable for its transactions.

Electronic Catalogs

Using this model, a supplier posts an electronic version of its catalog in a Web site for free access from interested parties. The company benefits from exposure to a large pool of potential buyers over the Internet without the costly creation and distribution of voluminous catalogs. The electronic catalog can be updated in a timely manner. Most companies have this model as a supplement ary to their paper-based catalogs to reach more customers outside their physical facilities. The transactions incurred may be handled with a traditional procurement process.

In this passive and low-cost business model, a supplier could inform potential buyers of the existence of the catalogs via regular mail or e-mail. The supplier may also register the Web site in the directories of some exchanges or intermediaries. Using a search engine, interested buyers may discover the competitive offer and then contact the supplier directly for further information about products and services.

Automated RFQs

In this model, requests for quotes (RFQ) are automatically distributed from the buying company to its business partners via a private communication network. An example of this model is GE's Trading Process Network (TPN). At GE, the sourcing department receives the requisitions electronically from other departments. It sends off RFQs containing specifications for the requisitions to a pool of approved suppliers via the Internet. Within a few hours, potential suppliers around the world are notified of incoming RFQs by e-mail or fax, instead of within days and weeks, as in the traditional paper-based system. Suppliers have a few days to prepare bids and to send them back over the extranet to GE. The bids are then routed over the intranet to the appropriate purchasing agents and a contract could be awarded on the same day. GE reports that, using TPN, labor involved in the procurement process was reduced by 30% and material cost was reduced from 5% to 50% due to reaching a wider base of supplier online. Procurement departments of GE's branches around the world can share information about their best suppliers. With TPN, it takes a few days for the whole procurement process instead of weeks, as before. Because the transactions are handled electronically, invoices are automatically reconciled with purchase orders and human errors in data entries/processing are minimized accordingly.

In this business model, sourcing cycle time in the acquisition process is reduced significantly, with the distribution of information and specifications to many business partners simultaneously. It allows purchasing agents to spend more time negotiating for the best deal and less time on administrative procedures. A company also consolidates a partnership with suppliers by buying only from approved sources and awarding business based on performance. Consequently, it allows the company to acquire quality goods and services from a large pool of competitive suppliers around the world.

With the advent of Web-based technology, networking becomes affordable and cost effective for interested businesses. A smaller company can engage an intermediary, or Web-service provider, to alleviate the cost of building and maintaining a sophisticated transaction network.

Digital Loyalty Networks

As in traditional business, highly valued business partners in B2B e-commerce may get special treatment. In this model, a B2B e-commerce Web site differentiates visitors by directing the valued ones to a special site, instead of trading in a public area opened for other regular business partners. The system may also direct special requests/offers to a preferred group of business partners.

Using this business model, different RFQs will be sent to different groups of potential suppliers from the approved supplying source for the company. A business may differentiate between its suppliers based on past performance in terms of product quality, pricing, delivery, and after-sales services. Similarly, a selling company may reward its preferred buyers with special discounts and conditions on transactions.

Metacatalogs

In this model, catalogs of approved suppliers are aggregated, indexed so that buyers will have the opportunity to deal with a large pool of suppliers of goods/services. These metacatalogs are usually kept in a central site for ease of access to potential buyers. Using this model, a global company may maintain a metacatalog of suppliers for the internal use of its branches. Or a trading mall can keep a metacatalog for the wide public access.

For the internal use of a global company, the model aggregates items of all approved suppliers from their catalogs into one source. Buyers from affiliated firms or branches can find the items in need, check their availability and delivery time, and complete an electronic requisition form and forward it to the selected supplier. In this transaction, prices could be negotiated in advance. Potential suppliers tend to offer competitive prices, as they would be exposed to a larger pool of buyers, in this case the world-wide affiliations/branches of the buying company. In addition, suppliers may become involved in a long-term relationship with a global company and its affiliations/branches. The listing in the metacatalog is free to the suppliers as a result of the negotiation of terms and prices for the goods/services to be provided to the buying company.

For wide public access, an intermediary or a distributor will create metacatalogs and make them available for its clients. Because buyers have an opportunity to deal with a large source of suppliers, these suppliers are under pressure to compete with one another in terms of price, quality, and services to win business. In this model, the supplier may have to pay a fee for listing on the catalog and/or a commission as a percentage of the transaction value. The buyer may have a free access or may pay a membership fee to the host/distributor.

Order Aggregation

In this model, RFQs from buyers are aggregated and sent to a pool of suppliers as invitations to tender. The order aggregation could be internal or external. In an internal aggregation, company-wide orders are aggregated to gain volume discounts and save administrative costs. In an external aggregation, a third party aggregates orders from small businesses and then negotiates with suppliers or conducts reverse auctions to reach a deal for the group. Usually, an intermediary will aggregate RFQs of participant buyers and match then with requests for proposals (RFPs) from participant suppliers.

In order aggregation, small buyers benefit from the volume discount through aggregation that could not be realized otherwise. Similarly, suppliers benefit from providing a large volume of goods/services to a pool of buyers and save the transaction costs incurred from dealing with many, fragmented buyers. Order aggregation works well, with defined indirect production materials and services having relative stable prices. In this model, if the order aggregation is undertaken by an intermediary, then involved business parties may have to pay a flat fee and/or a commission on the transaction value.

Auction

To reach a deal, business partners involved in B2B could use auction and/or matching mechanisms. A *forward auction* involves one seller and many potential buyers. A *reverse auction* involves one buyer and many potential sellers. In *double auction,* buyers and sellers bid and offer simultaneously. In *matching,* related price, quantity, quality, and delivery terms from the bid and ask are matched.

In a buying-side marketplace, a buyer opens an electronic market on its own server, lists items in need, and invites potential suppliers to bid. The trading mechanism is a *reverse auction,* in which suppliers compete with one another to offer the lowest price. The bidder who offers the lowest price wins the order from the buyer. Other issues, such as delivery, schedule, and related costs, are also taken into account when awarding contracts to bidders.

In a selling-side marketplace, a seller posts the information for the goods/services to be disposed and invites potential buyers to bid. The trading mechanism is a *forward auction,* in which participating buyers compete to offer the highest price to acquire goods/services in need.

The transaction can also take place at an intermediary site, at which buyers post their RFQs and suppliers post their RFPs. Depending on the regulations of the auction site, bidders can bid either only once or many times. In the latter case, bidders can view current supply and demand for the goods/services and change their bids accordingly. The transaction concludes when bidding prices and asking prices are matched.

The advantage of this model is that it attracts many buyers to a forward auction and many suppliers to a reverse auction.

The auction can be conducted at the seller/buyer private trading site or at an intermediary site. The auction can be in real time or last for a predetermined period.

If the auction is conducted at an intermediary site, the involved business parties may have to pay an access fee. In addition, sellers may have to pay a commission on transaction value.

Bartering

In this model, a company barters its inventory for goods/services in need by announcing its intention in a classified advertisement. Actually, a company rarely finds an exact match by itself. The company will have a better chance if it joins an e-commerce trading mall, as it could reach a larger pool of interested parties over the Internet.

An intermediary can create a bartering exchange, at which a company submit its surplus to the exchange and receives credits. Then it can use these credits to buy the items in need from the stock of goods/services listed for bartering at the exchange. An example is the bartering site of *Intagio.com* (formerly *Bartertrust.com*), where the owner claims to be the market leader in facilitating corporate trading in which goods and services are exchanged between businesses without using cash. Business parties using an intermediary site may have to pay for a membership fee and/or a commission on the transaction volume.

MERITS AND LIMITATIONS OF B2B E-COMMERCE

Along with other business models in the e-marketplace, B2B e-commerce has been welcomed as an innovative means of conducting transactions over the Internet. These business models promise not only effective and efficient business operations/transactions, but also competitive advantages to early adopters. Companies have adopted B2B e-commerce more slowly than predicted, but even conservative projections estimate that B2B transactions will top $3 trillion by 2004 (mro.com, 2002). It has been predicted that 66% of bids for MRO goods will be solicited over the Internet and 42% of MRO orders will be electronic (Fein & Pembroke Consulting, 2001).

An example of cost savings in B2B e-commerce is the story of Suncor Energy Inc. of Canada. In 2000, the company was working with 1,000 suppliers, with a total MRO budget of $192 million. About 70% of its expenditures went to 40 suppliers, yet its purchasing staff spent 80% of their time manually processing transactions with the 900 smallest suppliers. The switching to e-Procurement was predicted to generate a savings of $32 million directly from e-Procurement, $64 million from the redeployment of the purchasing workforce, and $10 million in inventory reductions (mro.com, 2002). It is noteworthy that these savings come from process automation, not from forcing price concessions from distributors.

Although having certain merits, these business models encounter some limitations that hinder the effective and efficient implementation and operation of a sustainable business. However, there are many possible solutions for overcoming these limitations.

Merits of B2B E-commerce

B2B e-commerce in general exposes a selling/buying company to a larger pool of suppliers and corporate buyers. Transactions over the Internet help overcome the geographical barrier, bringing business partners from all over the world to the e-marketplace. A company may benefit from transactions with business partners beyond the local market.

The Web-based technology of e-commerce helps minimize the human error found in the paper-based activities and supports timely, if not real-time, communication between and among partners. Different from traditional, costly telecommunications networks, Web-based technology makes transactions over the Internet affordable to most businesses involved in the e-marketplace. Also, the existence of many intermediaries also provides interested businesses with low-cost solutions for implementing a B2B e-commerce model.

B2B e-commerce models address the concerns about the effectiveness and efficiency of the supply chain management of business partners—suppliers as well as company buyers. Supply chain management coordinates business activities from order generation, order taking to order distribution of goods/services for individual as well as corporate customers (Kalakota & Whinston, 1997). Interdependencies in the supply chain create an extended boundary that goes far beyond an individual firm, so that individual firms can no longer maximize their own competitive advantage and therefore profit from cutting costs/prices. Material suppliers and distribution-channel partners, such as wholesalers, distributors, and retailers, all play important roles in supply chain management. B2B e-commerce models address the creation of partnerships with other parties along the supply chain, upstream as well as downstream, to share information of mutual benefit about the need of final customers. The key issue is that all upstream and downstream business activities should be coordinated to meet effectively the demand of final customers. Each partner in the stream should coordinate its own production/business plans (order fulfillment, procurement, production, and distribution) with those of the other partners so that sufficient streams of goods/services will reach customers in the right place at the right time.

B2B business models also address issues of customer relationship management (Kalakota & Whinston, 1997), the front-end function of a supply chain. An effective business model helps in creating more loyal customers who are not inclined to shop for lower prices but rather who pay for quality and service, in retaining valued customers, and in developing new customers by providing them with new quality products and services. The customer base could be segmented on history of performance in sales/purchases. This information will serve as a basis for promotion and discount, promoting the loyalty of current customers.

Limitations of B2B E-commerce and Possible Solutions

Some limitations of B2B e-commerce have been identified, such as conflicts with the existing distributing channel, cost/benefit justification for the venture, integration with business partners, and trust among business partners (Laudon & Traver, 2001; Turban et al., 2002)

Most suppliers have existing distributing networks of wholesalers, distributors, and dealers. If a company decides to do business over the Internet directly with interested partners, it may cause conflict in terms of territory agreement and pricing policies on product lines. A possible solution could be redirecting these potential customers to the appropriate distributors and having the company handle only new customers outside the current sales territories of these distributors. Another alternative could be the company handling specific products/services not available within the traditional distribution channel. Or orders could be taken at the central site, with a distributor providing downstream added-value services (delivery, maintenance, support) to the new customers of the company.

Another limitation is the number of potential business partners, and sales volume must be large enough to justify the implementation of a Web-based B2B system. Selling-side marketplaces for B2B e-commerce is promising if the supplier has a sufficient number of loyal business customers, if the product is well known, and if the price is not the critical purchasing criteria. For the buying side, the volume of transactions should be large enough to cover the investments and costs in the B2B e-commerce venture. In many cases, the interested business could participate in an exchange by paying a fixed fee or a commission on the volume of transactions. Using an intermediary could be feasible, as the company would not need to invest and maintain the expensive and sophisticated infrastructure of B2B e-commerce systems.

On a technical perspective, unless a B2B e-commerce site has implemented a comprehensive network/system architecture, integration with a variety of business partners systems (Oracle, IBM, or other ERP systems) may cause an operational problem. These business partners should be able to transact on compatible network platforms and protocols of communication. Sometimes the conversion implies additional investments and requires an extra cost/benefit analysis for the project. Also the technology should handle global transactions, such as multiple currencies and multiple languages from multiple countries, multiple terms of contract, and multiple product quality standards. *Commerce One* has been offering a "Global Trading Web" solution to address these issues.

Because transactions over Internet are not face-to-face, most business partners are unknown to each other. Consequently, the issue of trust in B2B is the same as in B2C e-commerce transactions. Many B2B exchanges have failed because they did not assure the creditability of the involved business partners. Trust in e-commerce could be enhanced with some quality assurance services and warranty seal programs, such as *WebTrust* and *SysTrust* of the American Institute of Certified Public Accountants (AICPA) (Nagel & Gray, 2001). In these programs, a third party (such as a CPA) audits the e-commerce transactions and infrastructure of a company to assure that it implements and follows some procedures and policies to guarantee the security of the online transactions and integrity in terms of fulfilling its obligation toward and honoring the privacy of its business partners. Once the company meets some prescribed criteria, it is awarded with a warranty seal to post on its Web site to inform the potential business partners on the security and quality of its online transactions.

CRITICAL SUCCESS FACTORS FOR B2B E-COMMERCE

From the performances of current B2B e-commerce entities, one can highlight some critical success factors having an impact on sustainable business and competitive advantages (Laudon & Traver, 2001; Turban et al., 2002).

A company has pressure to cut costs and expenses in the traditional paper-based procurement process related to vendor and product searches, vendor performance and cost comparison, opportunity costs, and errors of manual system. B2B e-commerce would provide ample opportunities and alternatives to optimize the procurement process. In this circumstance, the company has an incentive be involved in an effective and efficient cost-saving venture using Web-based technology. In addition, the top management will be interested in sponsoring and advocating the project.

Another success factor would be for a company to have experience with EDI or other non-Web-based business-to-business electronic transactions and be willing to integrate its current systems with new technologies in B2B. This would create a favorable climate supporting technology innovation. This factor is important in evaluating the technical feasibility of the B2B e-commerce project. It helps assess the readiness of the company, in terms of its technological maturity, to nurture an innovative system, and the availability of technical expertise needed to develop, operate, and maintain the system.

The industry concentrates on selling and buying with fragmented supplier and seller, and experiences difficulties in bringing both parties together. The larger source of buyers and sellers offered by B2B e-commerce would provide a company with opportunities to optimize its supply chain management. In this context, the potential economic and operational benefits would justify involvement in a B2B e-commerce venture.

Large initial liquidity is needed in terms of the number of buyers and sellers in the market and the volume and value of transaction to attract early business venture. In any economic feasibility analysis of a new venture, one needs to assess the cost of development and the payback period of the system. A large initial liquidity would justify the initial investment in a sustainable business.

A full range of services, such as credit verification, insurance, payment, and delivery, is needed to attract small and medium businesses. The market maker also needs available domain expertise for these services. The added-value services would facilitate the transactions of smaller businesses. These business partners, without a sophisticated infrastructure and expertise, will need a one-point access to the e-marketplace to conduct a one-stop transaction in B2B e-commerce.

Business ethics should be respected, to nurture trust among business partners, fairness to all business parties, especially in non-face-to-face transactions over the Internet. Security issues should be implemented to protect privacy and trade secrets of involved business entities in an open networked marketplace. To address the issue of trust, the company may include some quality assurance services and seal programs, such as WebTrust and SysTrust of AICPA.

Another factor is being able to successfully manage the channel conflict, to avoid any impact on the short-term revenue of supply chain partners. This conflict of interest is one of the limitations of B2B e-commerce and possible solutions to it were discussed in the previous section.

B2B E-COMMERCE ENABLE TECHNOLOGIES AND SERVICES

The hands-off nature of such Internet technologies as the communication protocols TCP/IP and HTTP and the programming languages HTML and XML enables the progress of e-commerce. These technologies assure interoperability across businesses using various platforms, which is necessary for global communications and transactions. XML, as a widely distributed standard, is compact and easy to program and permits businesses to more completely describe documents and transactions over the Internet.

Involvement in B2B e-commerce does not necessarily require intensive investment in hardware, software, or staffing for a sophisticated telecommunications network and database. There are many application and service providers that offer cost-efficient solutions for business interested in participation in this innovative marketplace. These services may support the complete value chain processes/activities of a business from its upstream suppliers to its final customers. Below are reviews of some service providers in B2B e-commerce.

GE's Global eXchange Services (GXS) (*gxs.com*) offers many services to B2B business. Its Trading Information Exchange (TIE) is a global extranet service with features such as online information publishing, dynamic delivery of supply chain data, and promotions–management workflow applications, enabling partners to share detailed operational information and to jointly manage key business processes. Its Source-to-Pay Services suite includes facilities for posting and responding to RFQs, online auctions, catalog purchases, invoice tracking, and payment. This provider has served more than 100,000 trading partners conducting about 1 billion transactions worth $1 trillion annually. GE itself has used these services internally since 1999 to connect with about 36,000 of its suppliers and has conducted about 27,000 auctions worth close to $10 billion (isourceonline.com, 2002).

Commerce One promotes the Global Trading Web to link buyers, suppliers, service providers, and e-marketplaces in a single, global community. This network is based on an open architecture—without closed and proprietary applications—and widely disseminated standards to assure technical and operational interoperability. The standards on data, content, and document format are established across the community. The technical interoperability assures that files, data, and applications can be transferred across platforms. The operational interoperability assures that e-marketplace processes and procedures can operate in unison. There are about 10,000 buyers and suppliers participating in the Global Trading Web. In addition nearly 100 e-marketplaces in all sizes and industries make up the membership that forms the backbone of the trading community (commerceone.com, 2000).

Ariba (*ariba.com*) products and services have enabled B2B e-commerce processes for more than 100 leading companies around the world, including over 40 of the Fortune 100, in diverse industries. Ariba products have been implemented on more than 3,750,000 desktops around the world (ariba.com, 2002). Web-based marketplaces powered by Ariba unite fragmented value chains operated by interdependent trading partners, bringing together buyers, suppliers, and service providers in Internet-speed trading communities. Ariba provides solutions for the rapid deployment and configuration of online procurement portals and automates end-to-end commerce processes, including catalog searches, requisitioning, purchasing, and invoicing. Integration with Ariba Supplier Network provides the infrastructure and third-party services companies need to transact, manage, and route orders in real time. It connects companies to e-procurement on-ramps, supplier-hosted catalogs, and other marketplaces. Ariba Marketplace integrates seamlessly and comprehensively with dynamic sourcing and RFQ capabilities, providing market makers with cost-efficient trading models, such as auctions, reverse auctions, bid/ask exchanges, and negotiations features. It can also handle real-time multicurrency translation for increased payment capabilities.

i2 (*i2.com*) has over $1 billion in revenues with more than 1000 customers. It has delivered about $30 billion in audited value to customers (i2.com, 2002). i2 Global Network is an Internet collaboration space that enables buyers, suppliers, and marketplaces to rapidly connect to each other and use i2 Network Services for content, collaboration, and commerce. These services allow the enterprise to extend its e-procurement and collaboration initiatives beyond tier 1 suppliers. i2 industry solutions include preconfigured industry templates, packaged role-based workflows, integration capabilities, product configurations, and example models and scenarios built specifically for an industry. From this starting point, companies can easily modify the template to meet their unique needs. i2 Supplier Relationship Management (SRM) supports the partnership with suppliers by coordinating processes across product development, sourcing, supply planning, and purchasing within a company and across companies. i2 Supply Chain Management (SCM) manages the supply chain within a company and across companies in the value chain as well. It provides multienterprise visibility, intelligent-decision support, and execution capability utilizing open, real-time collaboration with trading partners. i2 Demand Chain Management (DCM) synchronizes customer front-end processes with operations, enhancing responsiveness of the supply chain to maximize customer profitably and loyalty.

MRO Software Inc. (mro.com) provides fully hosted, off-the-shelf online services for Web storefront, security, catalog management, supplier administration, customer relationship management, order management, transaction processing, and integration with other online B2B services in the e-marketplace. This provider has served more than 8,000 customers (mro.com, 2002).

BEYOND SELLING AND BUYING B2B MODELS

B2B e-commerce extends to activities other than just selling and buying. For example, partners in a value chain could be involved in *collaborative commerce* (c-commerce) in a Web-based system to meet final consumer demand by sharing information on product design, production planning, and marketing coordination. Once consumer demand is identified, the quantity on hand of the raw material and semifinished and finished products of one partner will be made visible to others, avoiding bottlenecks along the value chain and supply chain (Laudon & Travers, 2001). In this type of business, some partners act as value chain integrators while others are value chain service providers. This business model assures the production of goods/services that effectively meet consumer demand with the collaboration between manufacturers and retailers. Then the product design and production cycle will be efficiently shortened with the collaboration between manufacturers and upstream suppliers.

GLOSSARY

Aggregation of orders and/or RFQs A compilation of small orders and RFQs of many businesses into a larger package to gain volume discount and economic of scale.

Bartering A trading method in which business partners exchange their surplus to one another without using cash.

Company-centric B2B e-commerce A business model represents a one-to-many business relationship in B2B e-commerce. In this model, a company involves in direct selling or direct buying of goods and services with many business partners.

Digital loyalty network A business model to offer special treatment to valued/preferred business parties in the value chain or supply chain in terms of priority, pricing and contract conditions.

Exchange/trading mall A business model represents a many-to-many business relationship in B2B e-commerce. In this marketplace, many buyers transact with many suppliers for goods and services.

Meta-catalog A compilation and index of goods and services offered by many small businesses into one source for easy of access to the public or interested parties.

MRO Non-production or indirect materials in maintenance, repairs, and operations.

Request for proposal (RFP) A tendering system in which a seller lists the materials for disposal and asks potential buyers bid on the contract. Buyer offers highest bid (forward auction) will win the contract.

Request for quote (RFQ) A tendering system in which the buyer lists the materials in need and asks the potential suppliers bid on the contract. Supplier offers lowest bid (reverse auction) will win the contract.

CROSS REFERENCES

See *Business-to-Business (B2B) Electronic Commerce; Business-to-Consumer (B2C) Internet Business Models; Click-and-Brick Electronic Commerce; Collaborative Commerce (C-commerce); Consumer-Oriented Electronic Commerce; Electronic Commerce and Electronic Business; E-marketplaces.*

REFERENCES

Ariba marketplace. Retrieved August 2002, from http://www.ariba.com

Conquering the catalog challenge: A white paper. Retrieved October 2001, from http://www.mro.com

A distributor's road map to e-business: A white paper. Retrieved February 2002, from http://www.mro.com

Fein, A. J., and Prembroke Consulting (2001). *Facing the forces of changes: Future scenarios for wholesale distribution*. Washington, DC: National Association of Wholesalers.

GE GXS enters the sourcing area. Retrieved May 2002, from http://www.isourceonline.com

The global trading Web—Creating the business Internet: A white paper. Retrieved November 2000, from http://www.commerceone.com

i2 corporate overview. Retrieved August 2002, from http://www.i2.com

Intelligent supply chain. Retrieved May 2002, from http://www.gegxs.com

Kalakota, R., & Whinston, A. B. (1997). *Electronic commerce: A manager's guide*. Reading, MA: Addison–Wesley.

Laudon, K. C., & Traver, C. G. (2001). *E-commerce: Business, technology, society*. Boston: Addison–Wesley.

Nagel, K. D., & Gray, G. L. (2001). *CPA's guide to e-business: Consulting and assurance services*. San Diego: Hartcourt Professional Publishing.

Pavlou, P. A., & El Sawy, O. A. (2002). A classification scheme for B2B exchanges and implications for interorganizational e-commerce. In M. Warkentin (Ed.), *Business to business electronic commerce* (pp. 1–21). Hershey, PA: Idea Group Publishing.

Trading information exchange electronically links retailers with their suppliers. Retrieved May 2002, from http://www.gxs.com

Turban, E. F., King, D., Lee, J., Warkentin, M., and Chung, H. M. (2002). *Electronic commerce 2002, a managerial perspective*. Englewood Cliffs, NJ: Prentice Hall International.

WebTrust principles and criteria. Retrieved January 2001, from http://www.aicpa.org/assurance/webtrust

Business-to-Consumer (B2C) Internet Business Models

Diane M. Hamilton, *Rowan University*

INTRODUCTION

For as long as commerce has existed, there have been diverse business models. A business model, very simply stated, is the method by which a firm manages to remain a going concern, that is, the way in which a company earns sufficient revenue to remain in business. By far the most often encountered business model is one in which an organization sells a product or provides a service in exchange for currency. Consider the *manufacturing* business model and, as an example, Dell Computer. Dell is in the business of building personal computers and selling them either to other businesses or to the general public. With the *retailing* business model companies also sell products in exchange for cash payment—for example Zales Jewelers. Zales does not manufacture the jewelry it sells; it provides a retail outlet, that is, a physical location, where potential customers can engage in the purchase of the goods that Zales has to offer. A variant on the retail model is the *catalog* business model. Catalog companies often don't have physical facilities (although some do); rather, they offer their goods for sale via their catalog and then ship the goods to the purchaser. An example of a catalog retailer is Figis, which sells gourmet snacks and gifts. The *service* business model has been growing significantly over the past several decades. Service-type businesses are quite varied, including, for example, beauty salons, attorneys, plumbers, and physicians. What they all have in common is that a service is rendered in exchange for a cash payment. Not all business models involve providing a product or service in exchange for cash. Consider, for example, the way some health clubs provide child care to their members. In some cases, the health club provides a room where children can remain happily occupied while their parent exercises. In order to avail themselves of the facilities for their children, health club members must be willing to personally provide the supervision for a fixed number of hours. That is, health club members who take advantage of the room by leaving their children there when they exercise pay for this benefit by working a few hours themselves in the child-care room. This service is paid for, then, not by cash but by a corresponding service. For any business model to be successful, it must generate some type of revenue or "value" that will enable the organization to continue in operation.

The arrival of the Internet's World Wide Web has brought about myriad new business models as well as a variation on the business models that had already existed, most notably the retailing model. These new business models take advantage of the Internet in many ways: to change a delivery system (e.g., digital delivery of software instead of physical shipment of the product); to improve customer service (e.g., online tracking of orders through the United Parcel Service); or to reach a wider audience (all geographical and time constraints are removed). Regardless of the model, however, the principle remains the same. A successful business model is one in which an exchange occurs between entities (companies or individuals) such that the organization can be self-sustaining through the receipt of revenue or something else of value, and it is worth noting that no level of sophisticated technology can make up for the lack of a good business model.

The next section is devoted to the definition and illustration of the most popular business-to-consumer (B2C) Internet business models. Later, some important tenets for Internet strategy are presented, along with lessons learned as a result of the quick rise and fall of many dot-coms.

INTERNET BUSINESS MODELS (B2C)

Some businesses have migrated to the Internet without changing their business model at all. These companies built Web sites, aptly called *brochureware,* which simply provide information about the company—as it exists in the physical world. They don't attempt to provide another sales channel, that is, to engage in virtual commerce. Other businesses, called brick-and-clicks, moved to the

Internet by offering their current products for sale online. Still others didn't exist prior to the advent of the World Wide Web and couldn't operate without it. There are clear advantages to electronic commerce from the perspective of both the business and the consumer. For example, an online business is available to a much larger set of customers than would be possible if the customer had to physically visit the business establishment. Also, this larger potential market allows a business to sell its product with lower marketing costs. Further, online businesses can actually improve customer service, for example through online help desks. Finally, an online business can better utilize human resources and warehouse/retail space.

Benefits also accrue to consumers who shop online. Customers can now shop at their convenience—any day and any time of day. They have a much larger selection to choose from; that is, it is possible to visit many more online retailers than would be feasible if they had to get in their car and drive from store to store. Price comparisons are more easily made online, especially with the aid of shopping bots, as is described later.

It is rare, however, to find a business environment that provides advantages without also having potential problems. Electronic commerce is no different. Problems are possible for both the online business and the online consumer. For example, online businesses suffer from a much higher rate of credit card fraud than their real-world counterparts, according to Visa—24% for online transactions compared with 6% overall for all transactions (Mearian, 2001).

In the following sections, the most popular business-to-consumer (B2C) business models are described and illustrated. It should be noted that although there are many ways to classify business models, no generally accepted categorization scheme for Internet business models currently exists.

The breakdown provided in this paper attempts to show the diversity of business models operating on the Web according to the type of transaction (or activity) engaged in, coupled with the way revenue is earned. Revenue can be earned, for example, through traditional sales purchases, as a commission on auction purchases, or through advertising that supports free information. Some of these business models, for example auctions and reverse auctions, existed in the physical world prior to the advent of the World Wide Web. However, the Web has allowed these business models to be redefined in virtual space, taking advantage of the huge Internet community. There are other business models that only came into existence after the advent of the World Wide Web, for example shopping bots. The way each of these unique models operates on the Internet is described in the following sections.

The Retail Model

Businesses adopting the *retail* model sell a product in exchange for cash. These retailers can be further differentiated according to whether the company exists solely online (online-only storefronts) or whether the Internet serves as just one of multiple sales channels (brick-and-click retailers). *Online-only storefronts* are more popularly called *dot-coms;* these businesses did not exist prior to the advent of the World Wide Web. As a matter of fact, it was the World Wide Web that provided the conditions allowing for the emergence of the dot-coms—companies that exist only on the Internet. Perhaps the most well-known dot-com company is Amazon.com, which started as a bookseller and migrated into a marketplace that now sells many diverse product lines, such as electronics, toys and games, music, cars, and magazine subscriptions. Amazon titles their home page "Amazon.com—Earth's Biggest Selection." They have clearly ascribed to the "reach" strategy as espoused by Evans and Wurster (1999) as they have grown their company. (This strategy, along with several others, is explained in the final section of this chapter.) Somewhat less well know than Amazon is Pets.com, one of the more famous "failures" in the sea of dot-coms, going out of business late in 2000. Amazon owned a large share in Pets.com. Unfortunately for Amazon, this was the second company they backed that went out of business (Living.com was the first).

Organizations conducting business in the physical as well as the virtual world are often referred to as *brick-and-click* companies (and sometimes *click-and-mortar*). In contrast to dot-coms, these firms had an established business before the advent of the World Wide Web, and after the Web was created, an online presence was added. More recently, firms in this category have found that the most successful strategy is to use the Internet as a way of supporting their primary channel—in most cases, the retail store. Although some companies may have created a Web site that's simply brochureware (information without any transaction ability), brick-and-clicks are capable of conducting retail business over the Web. Consider two extremely well-known, yet very different, brick-and click-companies, Lands' End and Wal-Mart.

Lands' End appeared on the Internet very early, in 1995, and today boasts the world's largest apparel Web site. Well known as a catalog retailer, Lands' End also sells through outlet stores and the Internet. One reason for their Internet success is likely their initial catalog experience. Catalog stores have much in common with electronic commerce and have, therefore, proved to be successful when the Internet channel is added. Another reason for Lands' End's success is the way in which they have utilized the Web to add value for their customers. For example, "My Virtual Model" allows customers to "try on" clothes after creating a virtual model of themselves by entering their critical measurements. Lands' End also provides a feature they call "My Personal Shopper," wherein customers answer a series of questions about themselves, thus allowing a virtual personal shopper to make various recommendations. Another one of their features, "Lands' End Custom," provides custom tailored apparel on request, and they also provide online chat facilities with customer service.

WalMart became the nation's number one retailer in the early 1990s and the nation's largest employer in 1997. It is famous for its huge superstores, where a consumer can find just about anything at an everyday low price. Walmart.com, a wholly owned subsidiary of Wal-Mart Stores, Inc., came into existence in January 2000, making it a somewhat late entry in the e-commerce market.

Wal-Mart has been successful on the Web because they have exploited their channels, rather than keeping them separate. For example, a consumer can make a purchase on the Web and return it to a store, if desired. Consumers can see the current Wal-Mart circular or the current advertised sales by visiting the Web site, and they can request e-mail notice of special values.

The Auction Model

The *auction* business model has existed for a very long time; consider, for example, Sotheby's, who auctions valuable items to the public in large halls where many prospective buyers can assemble. Although Sotheby's has now added online auctioning to their services, eBay has clearly revolutionized the auction business model. eBay calls itself "The World's Online Marketplace." Their mission is to "help practically anyone trade practically anything on earth" (http://www.ebay.com). eBay has reinvented the auction by including every conceivable type of item—from the most expensive (e.g., automobiles and rare collectibles) to the least expensive (e.g., junk souvenirs) to professional services (e.g., Web design or accounting). eBay is one of the very few dot-coms that managed to make a profit almost immediately after it appeared on the Internet. It earns revenue in several ways: from "posting" fees, from "special feature posting" fees, and from commissions on item sales. That is, sellers who wish to join the auction pay a very nominal posting fee (ranging from a few cents to a few dollars) in return for eBay providing space to show their item. If the seller wishes to use a special feature that might catch the attention of prospective buyers (e.g., highlighting an entry or listing an item at the "top" of a search), he or she pays another small fee. These two types of fees are paid by the seller whether or not the product eventually sells. However, because they are such nominal fees, there is very little at risk to the seller when the item is placed for bid. The final type of fee is collected by eBay if anyone bids on the item and it eventually sells. This fee is based on the amount of the sale; that is, the seller would pay more for a car that is auctioned than for a used book or CD that sells. What eBay provides is the *mechanism* for completing a transaction between a buyer and a seller. It provides the complete software interface for the transaction, and it does so splendidly. Its proven dependability and accuracy has attracted almost 40 million registered users and made it the leading online marketplace for the sale of goods and services—almost $15 billion was transacted on eBay in the year 2002. Many reasons exist for eBay's success, not the least of which is the flawless operation of its Web site, or the low cost for putting items up for bid. In addition to these basic items, however, eBay has provided added value in several other ways. For example, "Buy it Now" is a feature that allows a buyer to purchase an item immediately, at a stated price, if no one has yet placed a bid on the item, without waiting for the full auction time (typically a week) to expire. eBay even has a feature that allows bidders to participate (real time) in auctions that are currently being conducted in the world's leading auction houses. Other Web sites offer auctions (e.g., Yahoo and Amazon), but their volume and success is far less than that achieved by eBay since it came into existence, in 1995.

The Reverse Auction Model

Like the auction model, the *reverse auction* business model was not born on the Internet; businesses have utilized this approach for many years through requests for proposals (RFPs). Whereas an auction offers a good or service for sale and invites interested parties to bid higher and higher to obtain it, a reverse auction specifies the demand for some good or service and invites sellers to offer the item for sale at lower and lower prices. Although there are many reverse auction sites available on the Net, no individual Web site in the B2C market has become a household name. Examples of reverse auction sites include Respond and PillBid. Respond offers consumers the opportunity to request either a product or a service in widely diverse categories, such as automobiles, computers, legal services, travel, and home improvement. As a matter of fact, about 60% of all online requests are actually for services, rather than for hard goods, according to Respond. The most often requested service categories are hotels/motels, auto insurance, apartment/house rentals, DSL service providers, and day spas (Greenspan, 2002). Respond earns their revenue from two basic sources: affiliates and product/service providers. Any site can become an affiliate by adding a link to Respond's site, and affiliates are paid $1 every time a visitor clicks through and makes a request at Respond. Product/service providers can choose from one of two payment options: a subscription plan or a straight fee schedule. Under the subscription plan, providers are given access to all leads along with lead management and scheduling tools. Under the fee schedule, providers pay a small amount for each lead to which they choose to respond.

A niche-oriented reverse auction occurs at PillBid, where consumers post an interest in filling a prescription. PillBid serves as the intermediary between the consumer and pharmacies and offers an additional, value-added service: a search of the PillBot database. PillBot searches online pharmacies and advises the consumer about the lowest prices available on the Net.

What the Internet provides for any Web site utilizing this type of business model is the potential for a huge community of participants, which is important for the ultimate success of a reverse auction. The ability to facilitate auctions without human intervention and the potential for a very large number of participants are the factors provided by the Internet that allow for the sale of low-price items and the ability of a company to achieve profitability by charging very low commissions. This would not be possible in the physical world.

The "Name-Your-Price" Model

Many people consider Priceline to be a reverse auction, and it is often referred to as such in the literature. However, Priceline claims that it is not a reverse auction because sellers do not bid lower and lower for the ability to fill the demand for an item (www.priceline.com). Instead Priceline has patented a technology that they call *"Name Your Own Price."* Their system works as follows. Consumers make purchase offers with a fixed price for such things as airline tickets and hotel rooms. Additionally, they agree to some level of flexibility with respect to vendor, travel date, number of connecting flights, and so forth. Finally, they agree in advance to make the purchase (by supplying credit card information) if their price and flexibility levels can be matched by a vendor. Priceline then takes these offers and attempts to find a vendor that will provide the product or service for the terms specified. This business model exploits the Internet by improving information about supply to buyers and information about demand to sellers. That is, Priceline provides added value by serving as an information intermediary. Like other market makers, Priceline obtains its revenue as the difference between the price of acquiring the product or service and the price at which it turns the product or service over to the consumer (Mahadevan, 2000). Priceline, now a household name, has seen its share of success and failure. Its stock has traded at as high as $158 and as low as just a few dollars. It has found tremendous success in the sale of airline tickets; as a matter of fact, Priceline has claimed to be the largest seller of leisure air tickets in the United States (Anonymous, 2000). On the other hand, it lost so much money attempting to sell groceries and gasoline that it had to exit these markets soon after entry. This would lead one to conclude that some categories are far better suited to the name-your-price business model than others.

The Flea Market Model

Each of the three previous models (auction, reverse auction, and name your price) belong to the broad category of models categorized as the "broker" model by Rappa (2001). In its simplest form, a broker is a site that brings buyers and sellers together, and that facilitates buying transactions by providing a location and a system where the buying can take place. Like the previous models, the *flea market* model acts as a broker. However, it differs from the other models because the purchase transaction more clearly takes place between the buyer and the seller, rather than through some type of auctioning system. The interaction "feels" more like a typical flea market purchase. This type of model is employed at Half, which is a subsidiary of eBay. Sellers can post items for sale at Half, and when buyers search for them, they see the selling price and data about the current "condition" of all items available. Once the buyer selects an item, the sale is enacted immediately at the posted selling price. Half, acting as a broker, accepts payment from the buyer, deducts its net commission (based on selling price and shipping costs), and gives the proceeds to the seller. Half also earns revenue by inviting other Web sites to become affiliates (similar to the affiliate program explained earlier for Respond.)

An even more typical flea market transaction takes place at iOffer. iOffer has an interface that closely resembles eBay, but it is not an auction. At iOffer, a prospective buyer can search for an item and see a listing of items that match the search criteria. (If nothing matches the buyer's criteria, the buyer can post a free "want ad." These want ads are then used to provide relevant sellers with potential sales leads.) Once the buyer finds an interesting item, he/she can either buy the item at the selling price or make an offer. This second choice allows the buyer and seller to interact directly, making offers and counteroffers, until either the sale is transacted or the buyer "walks away." As with other broker sites, the seller pays a fee based upon the selling price of the item.

The Group-Buying Model

Manufacturers and distributors have always provided price breaks to buyers as the number of units ordered increased. That is, a company would expect to pay less per unit for an order of 1,000 units than for an order of one unit. The *group-buying* business model attempts to provide individual consumers with this same advantage. Operationally, items for sale appear on the group-buying Web site for a period of time, generally several days to a week. As time passes and more buyers join the "group" (i.e., place an order at the current price), the price goes down, reflecting the large volume discount. At the end of the buy cycle, all group members are billed for the final price. Group members who have placed an order are likely motivated to attempt to get others to join the group, causing the ultimate price to go down even further. This sounds like a good idea; however, in practice, this business model has failed in the B2C market. In January 2001, both of the well-known consumer group-buying sites, Mercata and MobShop, closed their doors to consumers. Mercata had performed well since their opening, in May 1999, and was the top-ranked online buying service in Spring 2000, according to Gomez Advisors. MobShop, which opened in October 1998, had been attempting to shift to the business-to-business (B2B) and government markets when it discontinued consumer service. MobShop's software is now used by the U.S. General Services Administration and some B2B marketplaces (Totty, 2001).

The Shopping Bot (Buyer Advocate) Model

A "bot" (short for robot) is a software tool that can search through tons of data on the Internet, return the found information, and store it in a database. Shopping bot programs crawl from Web site to Web site and provide the information that will ultimately populate the databases of search engines and shopping sites. Shopping bots can provide prices, shipping data, product availability, and other information about products available for sale online. This data is then aggregated into a database and is provided to a consumer when he or she is interested in making a purchase online. Most shopping bots were started by small, independent organizations and quickly sold to large companies for proprietary use on their site. For example, Amazon bought the technology supporting Junglee in 1998 and MySimon was purchased by Cnet in 2000. Shopping bot technology is incorporated into many sites; for example, DealTime, is incorporated into AOL's shopping section and that of other portals, such as Lycos and iWon. Initially, bots were often blocked by retail sites for fear of uncompetitive prices. However, bots are now considered advantageous because they drive consumer traffic to the site (buyers can "click through" to the desired retailer from the bot site.) As a matter of fact, bots now partner with shopping sites for revenue enhancement. Multiple revenue streams exist in this business model. For example, merchants can pay for advertising space on the bot site to promote special incentives and sales. Some bot sites, such as mySimon and DealTime, allow merchants to pay for a higher position in the retailer listing, and mySimon also obtains revenue for placing some products in a list of so-called "recommended" items (White, 2000). Other bot sites, rather than accept revenue for a higher place in the listing, charge all retailers simply to list on the site (Borzo, 2001). It will be interesting to watch the success of this business model, which earned quick popularity by exploiting a strategy of affiliation with the consumer, but which has systematically moved away from affiliation with the customer and closer to affiliation with the retailers.

The Full-Service Market-Making Model

Market makers don't sell products they personally produce. Rather, they bring together potential buyers and sellers and provide a virtual marketplace where all parties can discuss and engage in transactions. When the products and services are all related, a full-service site emerges. Revenue generally comes from two areas—program fees and advertising.

Autobytel, for example, improves the vehicle purchasing process by providing data for the initial research and comparison of models, the actual auto purchase, and the ultimate auto resale. It also supports car ownership by listing data about recommended service schedules and vehicle recalls, and it gives users the ability to schedule service and maintenance appointments online. Finally, the site provides featured articles and links to myriad auto parts and accessories retailers. Autobytel describes itself as an "Internet automotive marketing services company that helps retailers sell cars and manufacturers build brands through efficient marketing and customer relationship management tools and programs." Autobytel is the largest syndicated car buying content network, and its four branded Web sites received nine million unique visitors in just the fourth quarter of 2001 (http://www.autobytel.com). Like a good many other Internet business models, Autobytel earns revenue from a combination of sources, including dealer program fees, advertising, enterprise sales, and other products and services.

The Online Currency Model

Initially, *online currency* was designed to fill a desire for consumer privacy. That is, many thought that consumers might want to make online purchases where the detail wouldn't show up on a credit card statement. In other cases, e-cash was designed to enable payment of micropurchases. Qpass online currency, for example, can be used to make micropurchases (e.g., data on mutual funds) at Morningstar online. Currently, there are very few opportunities for micropurchases on the Net; however, as revenues from online advertising continue to decline, micropurchases may yet emerge successful for the purchase of small information items, such as news and medical information. Online currency has also been used to encourage surfers to visit particular sites, and their visits were rewarded with small amounts of online currency. This has not proved to be a successful business model, however. Cybergold, for example, paid consumers for various actions, such as reading ads, answering surveys, and registering at Web sites. Beenz acted similarly and attempted to fill a niche by providing e-cash in various international currencies. Cybergold and Beenz have both gone out of business. Flooz, another failed brand, was marketed as a gift certificate for Internet shopping. The model used for all these online currencies was to take a small percentage from transactions at their participating retailers, that is, the same as the credit card model. Paypal, a successful brand, with an initial public offering in early 2002, mainly facilitates eBay transactions and processes payments via major charge cards or direct checking account access. It now offers multiple currencies to facilitate global transactions. In some ways it appears that generic online currency is a solution still waiting for a problem.

The Free-Information Model

Free information of all types is ubiquitous on the Internet. For example, virtually all major news organizations maintain Web sites with regularly updated news items. When people don't have access to a newspaper or a television set (and even when they do), they can visit their favorite news site to find out what's happening in the world, what the weather is expected to be, and how their favorite sports team is faring. CNN, USA Today, and others use these sites to strengthen their brand and earn revenue through advertising.

News isn't the only type of information available on the Web, however. Diverse types of free information can be found, for example, at MapQuest, Nolo, and Encarta. MapQuest allows users to get driving directions between any two addresses in the United States—at no charge whatsoever. Nolo provides free advice about various legal issues, for example as related to creating a will or obtaining a divorce, and Encarta, provided by the Microsoft Network (MSN), is an online encyclopedia. These sites collect revenue in various ways. For example, MapQuest licenses its technology to thousands of business partners and provides advertising opportunities using banner ads and electronic coupons for hotels, restaurants, and the like, which target consumers as they travel. Nolo sells legal software and books on its site, and the free information that drives users to the site could also entice a user to make a purchase there. The Encarta site provides a great deal of information for free, but the full Encarta Reference Library is available for sale on the site, as are many other products. (This is the MSN, after all.) Just how long so much free information will remain free is unknown. In the future, these types of sites might be excellent candidates for micropurchases, should that concept ever become popular.

Search Engines

Search engines provide free information. However, they are unique enough and important enough to be discussed as a subcategory of the free-information model. Search engines serve as indexes to the World Wide Web. At a search engine site, people use keywords to indicate the type of information they are looking for, and the site returns matches considered relevant to the user's request. The data provided by a search engine comes from their database, which is continually updated. These databases are populated and updated through a continual search of the Web by programs called spiders. A number of search engines exist, for example Google, Lycos, and AltaVista. Because the spider code differs from search engine to search engine, as does the capacity of the database, the results one gets will also differ. Portals (discussed below) sometimes partner with a search engine in order to provide search facilities; for example, Comcast utilizes Google on its home page and Yahoo uses Google as its back-end search engine (Notess, 2002).

Directories

Directories are different from search engines in that a user typically selects from among a choice of categories, which have multiple levels of subcategories, to find the desired data. Yahoo was the first directory on the Web, and in addition to grabbing territory early, it has continually grown and reinvented itself. Yahoo now provides both directory and search engine facilities. It provides a tremendous amount of content, as well as services (e.g., e-mail), and shopping. As a result of the immense capability of its site, users spend increasingly large amounts of time there, and this, in turn, provides a great deal of information to Yahoo. This information is used to target advertising, which yields significant revenue, because targeted ads can sell for up to 60 times more than untargeted ones (Anonymous, 2001).

Portals

Although *portals* don't necessarily have to provide free information, in reality they generally do, so they are best discussed as a subcategory of the free-information model. A portal is a place of entry into the virtual world, that is, the site that a user chooses as his or her "home" in a browser, and most portals are, in fact, search engines or have search engine capabilities built in. Yahoo, for example, is one of the most popular sites used as a Web portal. Many other sites serve as the portal of choice for Web users, most notably those provided by Internet service providers, such as AOL or MSN. Because these general-purpose portals appear by default on the desktop when a browser is launched, they have a distinct competitive advantage. News sites such as USA Today or CNN are also popular portals. General-purpose portals typically provide current news items, shopping, and the ability to configure the page according to personal interest. Many also provide e-mail capabilities, online calendars, and Web-based storage. Some Web users prefer a portal targeted to their specific interests, such as Healtheon/WebMD, funtrivia.com, or MavenSearch (a portal to the Jewish world). Regardless of the type, portals derive their revenue from site advertising and partnership agreements, or in the case of Internet service providers (discussed below), subscription fees help to defray the costs of maintaining their portal.

The Service-for-a-Fee Model

Although products can and are purchased on the Web, services are available for online purchase as well. The *service-for-a-fee* model operates in the same fashion as the retail model in that consumers pay for services rendered, and these fees account for the great majority of earned revenue. At Resume.com, for example, professionals will write, revise, or critique a resume. They will also prepare cover letters and thank you letters or even serve as a "personal career agent." The price for these services ranges from $50 for a cover letter, to $99 for a "classic" resume, to $1,499 to act as a personal career agent.

In the same basic industry, Monster provides services to both job seekers and employers. Monster's slogan is "Work. Life. Possibilities." They accept job postings from both employers and potential employees and allow either group to browse the available listings. The price for a job posting starts at $120, which is much less than many large, urban newspapers will charge, yet the audience reach is significantly broader. They also provide a number of job-related services, such as the "Personal Salary Report," which gives job seekers an idea of their potential worth in the current job market, and the "Inside Scoop," which purports to tell job seekers what interviewers "really want to know." Monster is the largest employment site on the Internet, operating in at least 21 countries, thus providing a huge marketplace for employers and employees to meet. They are a real success story among the dot-coms, having about 30 million unique visits and over 12 million resumes. Yet they continue to seek avenues for new growth. For example, Monster recently joined with AOL, providing it with the largest consumer base on the Net, over 30 million members (www.tmpw.com), and it began providing online training and employee development services in August 2001 (Moore, 2001). This site is especially useful for employees who are willing to relocate or for employers who want the broadest possible geographical exposure for their positions. A search function using keywords is available for both employers and employees, to simplify the search process.

E*Trade, an online brokerage and banking firm, was launched in 1983 as a service bureau before the World Wide Web came into existence. It linked with AOL and CompuServe in 1992, becoming Etrade.com, one of the first all-electronic brokerage firms. It is now a global leader in personal financial services, providing portfolio tracking, free real-time stock quotes, market news, and research. E*Trade Bank, now the largest purely online bank, was added to its portfolio in 2000 to offer a variety of financial services, including checking, saving and money market accounts, online bill paying, mortgage and auto loans, and credit cards. In 2001, E*Trade announced its first proprietary mutual fund, which is to be solely managed by E*Trade Asset Management (www.etrade.com).

Internet Service Providers (ISPs)

Internet service providers are a little different from other types of services available on the Web, because they are not generally thought of as a "destination" on the Internet. Rather, ISPs provide consumers with the ability to access the Internet, and to send and receive electronic mail. Some of the most well known of these ISPs are AOL, CompuServe, and MSN. More recently, cable providers (e.g., Comcast and Cox) and DSL providers (e.g., Verizon) have earned and continue to increase market share by providing a high-speed alternative to modem access. As of early 2002, 7% of U.S. households were using one of these broadband alternatives for Internet access and the percentage is growing (Associated Press, 2002). ISPs charge a monthly fee for their service, which accounts for the vast majority of their revenue.

NetZero is the most well-known *free* ISP, but in return for being "free," it requires users to allow their surfing and buying behavior to be snooped. Thus, although the user is not paying for this service with cash, he or she is paying for it by providing a large amount of personal information. This information is worth a great deal to NetZero, who can turn it into cash by selling the information to other businesses. NetZero also fills the user's screen with more than the usual amount of advertising banners. Some believe that the free ISP model is not sustainable in the long run (Addison, 2001), and it should be noted that even NetZero has two types of service—free and "platinum." Platinum costs less than most well-known ISPs (currently $9.95/month), but it offers better service than the free subscription does, and it removes the plethora of banner ads from the screen. Other examples of the free-service model are mentioned below. NetZero is an example of both the service-for-a-fee model and the free-service model. It is included here because it is an ISP.

The Free-Service Model

It is often difficult to distinguish between free information and free services. For example, are the driving directions Mapquest provides a service or information? For purposes of categorization in this chapter, a *service* that actually provides *information* is considered as belonging to the free-information model. Only services providing something distinctly different from information are categorized as service models. One type of *free service* is the provision of games, such as games of chance. At Freelotto, for example, visitors can play an online lottery, which is supported by online advertisements. Napster provided a much publicized example of an extremely popular, free-service site. At Napster, users could trade music files; however, in reality, no trading was actually required. Members provided music files free to anyone who wanted to download them, and they could download as many files as they desired. Napster was shut down as a result of a lawsuit brought by music providers. There are similar sites still operating (e.g., www.imesh.com), but none of these sites is anywhere as successful as Napster. It is worth noting that the music industry has attempted to replace the popular Napster service by offering "fee-for-download" sites, such as RealOne and Pressplay. However, these sites are relatively expensive (about $10/month) and have alliances with only a subset of the large record companies, resulting in members being able to download only music sold by alliance companies. Most sites that offer a free service earn their revenue through site advertising and/or reselling of consumer information.

INTERNET STRATEGY

Creation of a successful Internet business model must also include the clarification of a firm's strategy, including the consideration of how to best exploit the features of the Internet. It should be noted that Internet strategy is not the same as a business model. The strategy outlines the specifics of how a company will implement a given model, and any business model can be implemented by an almost unlimited variety of strategies. The fact that some companies succeed when using a particular business model (e.g., auction) while others fail clearly attests to this fact. During the process of delineating an Internet strategy, one must clearly define how the day-to-day operations of the organization will allow the firm to become self-sustaining within a reasonable period of time. The amazing proliferation of dot-coms, and the incredible losses that accrued to some, provide us with many "lessons learned," which are also briefly discussed in this section.

Competitive Advantage Through the Internet

Evans and Wurster (1999) suggest that success on the Internet will accrue to those organizations that exploit one of the following: reach, affiliation, or richness. *Reach* is defined along two dimensions: (a) how many customers a firm can reach and/or (b) how many products/services can be provided to a customer. The advantage of reaching a large potential customer base is obvious, and that's one of the main reasons why many businesses set up shop on the Internet. Customers from virtually anywhere in the world can shop online 24 hours a day, 7 days a week. On the Internet a business can also provide a huge number of products, significantly more than can be provided in a finite physical space. For example, Amazon, one of the Internet's pioneers, offers 25 times more books for sale than the largest bookstore anywhere in the world.

Competitive advantage can also be achieved through *affiliation*—specifically, through affiliation with the consumer. Consumers will have more trust in an organization that appears to be on "their side." For example, if someone is going to buy an automobile, whose opinion will carry more weight with the buyer, the auto manufacturer (e.g., Ford or Totota) or an unbiased third-party (e.g., Edmunds.com or *Consumer Reports*)? The Internet has seen the rise of new types of business models that affiliate themselves with the consumer, for example search engines such as Yahoo or Google. Consumers can visit these sites, request data, have their requests honored, and pay nothing for the service.

The final dimension suggested by Evans and Wurster is *richness*. They define richness as (a) how much detail the organization provides to the consumer or (b) how much information about the consumer is collected by the organization. In the first case, firms can achieve competitive advantage by providing significant detail to the consumer—more than the amount provided by their competition. This strategy will be most effective when the product is technical and/or changes frequently. Examples of the types of products that might benefit from a strategy of richness are computers, wireless phones, and digital cameras. A firm can also achieve competitive advantage if it holds a significant amount of information about its customers, because information is an asset when used effectively. Many online companies use information about their customers' buying habits to customize the site for each visitor and also to provide suggestions about additional products the visitor might be interested in.

Lessons Learned

The advent of the World Wide Web brought with it a rush of businesses hoping to be first to achieve market share online. During this frenzy of the mid-1990s, venture capitalists were willing to invest in almost any type of online organization. The belief seemed to be that just about any idea could bring a fortune online. Nothing appeared to be too extreme. However, by the late 1990s, many of the companies that, years later were still losing money, started to appear less elegant, and much of the venture capital funding that was so easily obtained earlier was drying up. One phenomenon worth considering is the reliance on Web advertising for revenue. In the early days of Internet surfing, the banner ad was something new and intriguing, and many people were motivated to click through, hoping to discover a "terrific deal" or to find out what they "won." However, this initial curiosity was followed by consumer disinterest, as a result of both previous disappointments and the sheer numbers of ads popping up everywhere. Business models that relied solely on advertising revenue were among those fighting for their virtual life. Considering history and the current environment, a variety of recipes for online success have been espoused. Patton (2001) suggests several rules for online success. One is *be diverse*. Travelocity, for example, has found success by moving from its initial business model, which was to sell airline tickets and collect revenue from online advertising, to its revised model, which no longer depends so much on advertising, relying instead on its extensive customer database to sell things other than airline tickets (e.g., hotel rooms, membership in travel clubs, and suitcases) This is also an example of using "richness" to achieve competitive advantage. A second rule is *exploit channels*. Retailers have historically used multiple channels—such as retail stores and catalogs (e.g., Victoria's Secret or L.L. Bean)—for selling their merchandise. The mistake made by some retailers was to consider the Web as a separate, competing channel, so that what a person bought in one channel, for example online, could not be returned to a different channel, such as the retail store. This practice was met with consumer frustration and dissatisfaction. On the other hand, the more successful enterprises exploited their channels, using one channel to promote another and treating all channels equally. This practice strengthened their brands and increased customer loyalty.

Hamel (2001) suggests that there are three ways to imminent online failure, which he labels "dumb," "dumber," and "dumbest." He claims that it is a *dumb* idea to miscalculate timing. That is, there are circumstances that require speed, such as where customer benefits are substantial and competitors are likely to appear quickly. On the other hand, there are circumstances that call for a slower approach, that is, where complementary products or new customer behaviors are required in order to take advantage of the product. Even *dumber* is to overpay for market share. For example, Pets.com paid $180 for every customer it ultimately acquired. This huge cost, plus its inability to differentiate itself, led to its speedy failure (Patton, 2001). In order to be effective, a company must acquire customers at a discount, not at a premium. The *dumbest* idea, according to Hamel, is to come to market without a good business model, and he claims that there are two fundamental flaws that will kill a business model. They are (a) misreading the customer (are there really customers who actually want what you are going to sell?) and (b) unsound economics (are there really enough customers available to render your business profitable?).

CONCLUSION

A business model is the method by which a firm manages to remain a going concern, and a business strategy helps to define the goals of the particular business model chosen. With the advent of the Internet, some existing business models were revised and other new models were invented. Many of the revised business models simply incorporated the addition of a new sales channel—the World Wide Web. Companies utilizing these models are called brick-and-clicks. However, simply adding this new sales channel has provided no guarantee of success. Rather, companies who added this channel successfully did so through their specific implementation strategy. For example, Lands' End added value to their Internet channel by incorporating special features and services that would entice people to shop on their Web site, and Wal-Mart utilized their Web site to strengthen their primary channel—the Wal-Mart store. These Internet strategies (i.e., value-added features or services and channel exploitation), along with others described above, helped aspiring brick-and-click companies achieve Internet success.

Some companies revised an existing business model by adding value that could only be provided by the Internet. For example, eBay reinvented the auction model by providing an efficient and effective online marketplace, where sellers could market their goods with very little risk and for a very reasonable expenditure level. This has resulted in a volume so large as to exploit both the reach and affiliation strategies as espoused by Evans and Wurster (1999). Another example, the free-information model, is sometimes a variation on the traditional broadcast (television or radio) model. Companies such as CNN and USA Today provide free information to the public, and they earn their revenue through advertising. In other cases, the free-information model more closely resembles an ongoing "promotion" using free samples. Consider, for example, the information provided on the Nolo site, which might entice the user to purchase the related software package.

Finally, the Internet has provided the catalyst for some completely new models, and these models would not exist without the Internet. Examples in this category are search engines, Internet service providers, and portals. These brand new business models exist for the purpose of helping people effectively access the World Wide Web. That is, ISPs provide the technology necessary to connect to the Web, portals provide a place to start activity once access to the Web is gained, and search engines help users find the right path to Web sites of interest. Portals and search engines exploit an affiliation strategy, and the most successful ones provide the quality and quantity of information and services desired by the consumer, thereby exploiting richness.

Regardless of whether a business model merely includes the Internet as a separate sales channel or adds value to an existing model in order to exploit the unique features of the Internet, or whether a model was invented specifically for the Internet, it has been shown that no firm can be successful without a good business model coupled with sound strategy, resulting in sufficient revenue to allow the firm to remain a going concern. It is also clear that technology alone, no matter how sophisticated it is, cannot overcome the problems inherent in a business model that, for example, misreads the customer or for which insufficient customers actually exist. The days of unlimited venture capital are over; future funding will require sensible models united with a profitable strategy.

GLOSSARY

Bot Programs that can autonomously search through data on the Internet and return the data for storage in a database (short for robot; also called *spider*).

Brick-and-Click Companies conducting business in both the physical and the virtual world.

Business Model Method by which a firm manages to remain a going concern.

Click-and-Mortar Companies conducting business in both the physical and the virtual world.

Dot-com Companies that exist solely on the Internet.

Ecash Special-purpose currency that can be spent when shopping online.

Internet Service Provider (ISP) Provides Internet services such as access to the World Wide Web and e-mail.

Micropurchase An online purchase having a cost of from just a few cents to a dollar or two.

Portal A place of entry into the virtual world; the site a user selects as "home" in a browser.

Search Engine An index to the World Wide Web; allows users to enter keywords that help to find desired information.

Spider Programs that can crawl from Web site to Web site and retrieve data that is stored in the database of a search engine.

CROSS REFERENCES

See *Business-to-Business (B2B) Electronic Commerce; Business-to-Business (B2B) Internet Business Models; Click-and-Brick Electronic Commerce; Collaborative Commerce (C-commerce); Consumer-Oriented Electronic Commerce; Electronic Commerce and Electronic Business; E-marketplaces.*

REFERENCES

Addison, D. (2001). Free web access business model is unsustainable in the long term. *Marketing,* August 9, p. 9.

Amazon.com (0000). Retrieved July 26, 2002, from http://www.amazon.com

Anonymous (2000). E-commerce: In the great web bazaar. *The Economist, 354*(8159), S40–S44.

Anonymous (2001). Internet pioneers: We have lift-off. *The Economist, 358*(8207), 69–72.

Associated Press. FCC report: *High-speed net access is growing*. Retrieved February 8, 2002, from http://www.hollywoodreporter.com / hollywoodreporter / convergence / article_display.jsp?vnu_content_id=1320918

Autobybel. Retrieved April 2, 2002, from http://www.autobytel.com

Borzo, J. (2001). A consumer's report—Searching: Out of order?—You may think you're getting the cheapest merchant when you use a shopping bot; but instead, you may just be getting the biggest advertiser. *Wall Street Journal,* September 24, p. R13.

Ebay. Retrieved March 15, 2002, from http://www.ebay.com

Encarta. Retrieved March 20, 2002, from http://encarta.msn.com

Etrade. Retrieved February 21, 2002, from http://www.etrade.com

Evans, P., & Wurster, T. (1999). Getting real about virtual commerce. *Harvard Business Review, 77*(6), 84–94.

FreeLotto. Retrieved March 19, 2000, from http://www.freelotto.com

Greenspan, R. (2002). *Service seekers stay in the 'hood. Ecommerce-guide.com.* Retrieved July 21, 2002, from http://ecommerce.internet.com/news/insights/trends/article/0,,10417_1135801,00.html

Half. Retrieved July 21, 2002, from http://www.half.com

Hamel, G. (2001). Smart mover, dumb mover. *Fortune, 144*(4), 191–195.

Imesh. Retrieved March 18, 2002, from http://www.imesh.com

Ioffer. Retrieved July 21, 2002, from http://ioffer.com

Lands' End. Retrieved March 17, 2002, from http://www.landsend.com

Mahadevan, B. (2000). Business models for internet-based e-commerce: An anatomy. *California Management Review, 42*(4), 55–69.

MapQuest. Retrieved March 12, 2002, from http://www.mapquest.com

Mearian, L. (2001). *Visa purchases online security on merchants, banks*. Retrieved July 21, 2002, from http://www.cnn.com/2001/TECH/industry/05/15/visa.security.idg/?related

Monster. Retrieved March 22, 2002, from http://www.monster.com

Moore, C. (2001). *IBM, Monster.com kick start corporate e-learning*. Retrieved December 29, 2001, from http://iwsun4.infoworld.com/articles/hn/xml/01/08/20/010820 hnelearn.xml

Napster. Retrieved March 17, 2002, from http://www.napster.com

Nolo. Retrieved March 19, 2002, from http://www.nolo.com

Notess, G. R. *Review of Google*. Retrieved March 19, 2002, from www.searchengineshowdown.com/features/google/index.shtml

Patton, S. (2001). What works on the Web. *CIO Magazine, 14*(23), 90–96.

Paypal. Retrieved July 26, 2002, from http://www.paypal.com

Pill Bid. Retrieved March 18, 2002, from http://www.pillbid.com

Pressplay. Retrieved March 14, 2002, from http://www.pressplay.com

Priceline. Retrieved March 21, 2002, from http://www.priceline.com

Qpass. Retrieved July 26, 2002, from http://member.qpass.com/MACHelpCenter.asp?ReturnUrl = %2Fmacwelcome%2Easp&BrandingID = 0

Rappa, M. *Business models on the web*. Retrieved December 12, 2001, from http://digitalenterprise.org/models/models.html

RealOne. Retrieved March 15, 2002, from http://www.realone.com

Respond. Retrieved July 21, 2002, from http://respond.com

Resume. Retrieved March 21, 2002, from http://www.resume.com

TMP Worldwide. Retrieved March 14, 2002, from http://www.tmpw.com

Totty, M. (2001). Openers—Changing clients: How some e-tailers remade themselves as B-to-B businesses. *Wall Street Journal,* May 21, p. R6.

Travelocity. Retrieved July 26, 2002, from http://www.travelocity.com

Wal-Mart. Retrieved March 22, 2002, from http://www.walmart.com

White, E. (2000). The lessons we've learned—Comparison shopping: No comparison—Shopping 'bots' were supposed to unleash brutal price wars; why haven't they? *Wall Street Journal,* October 23, p. R18.

Yahoo. Retrieved July 26, 2002, from http://www.yahoo.com

C

Capacity Planning for Web Services

Robert Oshana, *Southern Methodist University*

INTRODUCTION

The Internet is continuing to grow rapidly throughout most of the world. The most popular Internet applications continue to be the World Wide Web (Web), electronic mail, and news. These applications use a client–server paradigm: Systems called servers provide services to clients. Client–server describes the relationship between two computer programs in which one program, the client, makes a service request from another program, the server, which fulfills the request. Although the client–server idea can be used by programs within a single computer, it is a more important idea in a network. In a network, the client–server model provides a convenient way to interconnect programs that are distributed efficiently across different locations. The client–server model has become one of the central ideas of network computing. Most business applications written today use the client–server model. So does the Internet's main program, TCP/IP (transmission control protocol/Internet protocol). It is critical that these servers are able to deliver high performance in terms of throughput (requests per second) and response time (time taken for one request). As new capabilities and services make their way to the Web, the ability for forecast the performance of integrated information technology networks will be critical to the future success of businesses that provide and rely on these capabilities and services.

The explosive growth of Internet sites and e-commerce has presented new challenges in managing performance and capacity. Also, the ability to estimate the future performance of a large and complex distributed software system at design time can significantly reduce overall software cost and risk. These challenges have led to the development of a discipline of managing system resources called capacity planning, which Menasce and Almeida (2000, p. 186) described as "the process of predicting when future load levels will saturate the system and of determining the most cost-effective way of delaying system saturation as much as possible."

Performance management is an important part of the capacity planning process. The goal of performance management is to proactively manage current system performance parameters such as throughput and latency to provide the user adequate response time as well as minimal downtime.

Planning Considerations

Determining future load levels can be difficult if a disciplined process is not used. There have been many examples of systems that suffered economic and reputation damage as a result of not planning correctly and adequately for growth and capacity. Building systems "on Internet time" appears to have been a way to get around disciplined system development processes in some cases.

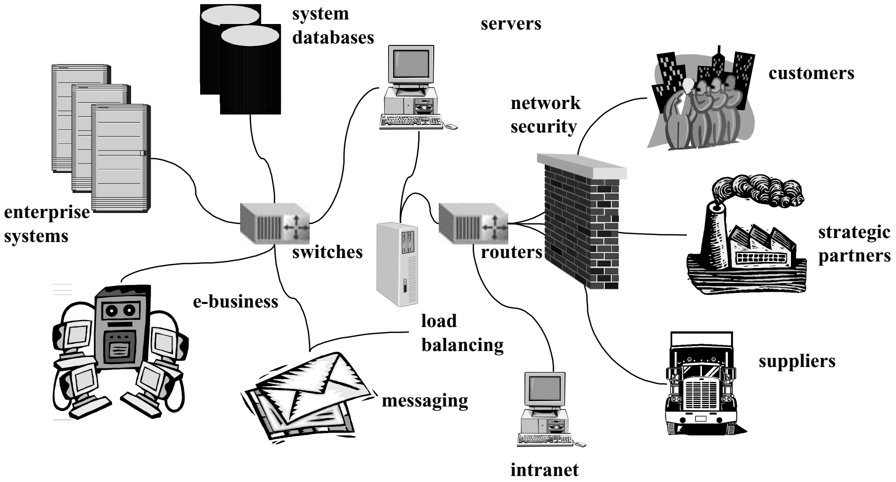


Figure 1: Modern information technology systems are complex networks with many stakeholders and performance requirements.

Modern information technology systems are complicated and involve a number of stakeholders and users (Figure 1). New capabilities are continuously being added to further complicate these already taxed systems. Planning the growth and evolution of these systems and managing them effectively cannot be a part-time job. There are too many interrelated factors that prevent it from being such, and businesses have come to be extremely dependent on these systems for their livelihoods.

Nevertheless, the determination of future load levels should consider how the workload for the system will evolve over time. If a Web system is just coming online, supplemented by an increasing advertising campaign, developers should expect an increasing workload due to an increasing number of visits to the site. If there are plans to deploy a system in increments, adding new capabilities periodically during a phased deployment schedule, this should be considered in the planning of future load levels. Finally, changes in customer behavior should be considered. Whether this is due to a sale, a world event, or seasonal functions being added to the site, the estimation should be predicted and managed accordingly. Regardless of the driver for future load level changes, the goal should be to head toward a predictive pattern and not based solely on experimentation.

Service Level Agreements

Service level agreements (SLAs; Figure 2) should be put in place with the customer (internal or external) to establish the upper and lower bounds of performance and availability metrics. For example, the customer may need to have the site availability greater than 99.95% for security reasons. This should be stated in the SLA. If the server side response time to information requests must be less than 5 seconds per request, this should also be stated in the SLA.

SLAs vary by organization. The creation of an SLA is a management function with feedback from the user as well as the information technology developers. If any of the IT functions are outsourced, the SLA becomes more important and should be the basis for subcontractor management.

Determining Future Load Levels

Determining future load levels of a system requires careful consideration of three major factors. The first is the expected growth of the existing system workload. Consideration should be given to how the company or business will evolve over time.

The second is the plan for deploying new system services and new or upgraded applications. These new capabilities will require more memory, processing power, and I/O (input/output). These capabilities must also be measured against how often they will be used in order to determine the overall system impact when they are deployed. For example, for a new semiconductor design workbench capability added to its Web site, an estimate of the processing resources required to run the new application as well as the number of new users drawn to the site that will be using the new capability must all be considered.

Finally, the third factor to consider is changes in customer behavior. This includes surges in site traffic due to world events, news stories, sales and advertisements, and other events that draw people to a specific site. The site must be able to sustain these temporal changes to load levels. If the semiconductor design workbench capability is accompanied by a large advertising campaign, the system should be built to handle the surge in new users registering for this capability, each of which will need system resources to operate their virtual design workbench.

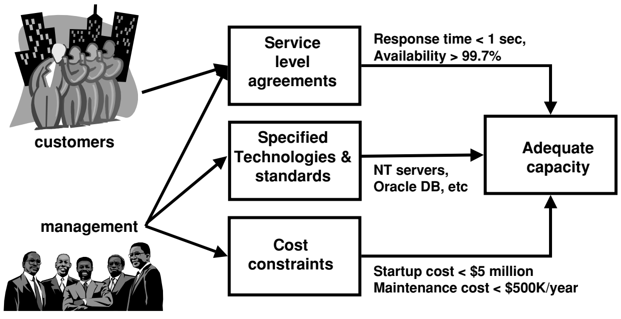


Figure 2: Capacity planning is driven by several system parameters.

In determining adequate capacity for a site, there are several factors that must be considered (Figure 2; Menasce & Almeida, 2000, p. 187):

- **System performance.** This includes factors such as performance and availability metrics for the system. Examples of performance metrics to be considered are server-side response time and session throughput. (A session in this context will normally consist of several Hypertext Transfer Protocol [HTTP] requests performed by a user during the period the user is browsing, shopping, etc.) Examples of availability metrics include site availability. Site availability is driven by many other factors including backup time, time to install new hardware and software, reboot time for failed applications, system maintenance time, and so on.
- **System technologies.** Internet solutions can draw from a number of technologies in the industry. Servers are available from many vendors that have provided certain advantages for specific types of applications (for example, transaction-based applications). There are many database solutions available as well. The choice of these technologies and the decisions of how to integrate these technologies has a significant effect on the overall site performance.
- **System cost.** System performance and system technologies are ultimately affected by the cost constraints for the system. Cost drives many of the decisions with respect to the type of servers, the number of servers, the technology used, training on the new technology, and so on.

CAPACITY PLANNING METHODOLOGY

Menasce and Almeida (2000) described a capacity planning methodology as consisting of four core planning processes: business, functional, resource, and customer behavior planning.

Business Planning

The main goal of this process is to generate a business model that describes the type of business conducted. Web-based businesses can be business to business (B2B), which is a business that sells products or services to another business; business to consumer (B2C), which is a business that sells products or services to a targeted set of end user consumers; and consumer to consumer (C2C), in which untrusted parties sell products and services to one another. Business planning also considers the type of delivery associated with the business model. For example, selling books requires a fulfillment model that is different from distributing an upgrade to a software application that may be a digital delivery directly over the Internet. Business planning must also consider the use of third-party services that can vary from such functions as fulfillment and delivery, to site maintenance, to customer support. Other quantitative measures are also considered, such as the number of registered users, which may be important for an online subscription product to consumer buying patterns, which may be important to the advertisers sponsoring some or all of the site development.

Functional Planning

A functional model of the system can be created that includes important information about the functions provided by the system including how the user interacts with the system, the Web technology used, the type of authentication used for e-business sites, and so on. The functional model is required before resource allocation can be estimated. The functional model can vary by business type and needs and will have a considerable impact on the overall resource allocation. The selection of the search engine, for example, will have a direct impact on the processing resources required to execute the particular search algorithms (which vary from search engine to search engine). The type of database model will dictate the type and form of query required, which will also drive resource demands. An online workbench for designers will require a certain amount of processing resources and memory to support this type of activity.

Resource Planning

Resource planning is used to map resources based on customer behavior models the functional and business planning. Resource planning requires the development staff to characterize the current information technology (IT) environment and model it for performance analysis. This model is then analyzed and calibrated, if necessary, iteratively until the model is validated against the requirements of the system and SLA. Because few organizations have infinite cost resources to accommodate all wishes, the performance model must be consistent with the cost model for the project. The iterative nature of this phase attempts to develop a system, as closely as possible, that matches both the performance model as well as the cost model of the system. Because of this goal, a lot of "what if" analysis is usually performed in this phase to achieve the most optimal solution.

Customer Behavior Planning

Customer behavior models are also required to perform adequate capacity planning. These models are used to determine the possible navigation patterns by the user. This information is then used to determine required or needed site layout adjustments as well as new services. For example, if the navigation pattern data showed users spending a majority of time performing information queries, the database server will need to be designed to accommodate this pattern. If the navigation patterns showed customers navigating to a function that displayed various product images, this information can be used to provide adequate performance in image presentation. The customer behavior characterization leads to workload characterization and forecasting. Based on the navigation patterns of the user and where the user spends most of the time when navigating a Web site, the workload can be estimated to provide a goal for resource allocation.

A well-thought-out and planned methodology for capacity planning will lead to more effective systems integration and execution. Effective capacity planning also depends on accurate predictive performance models.

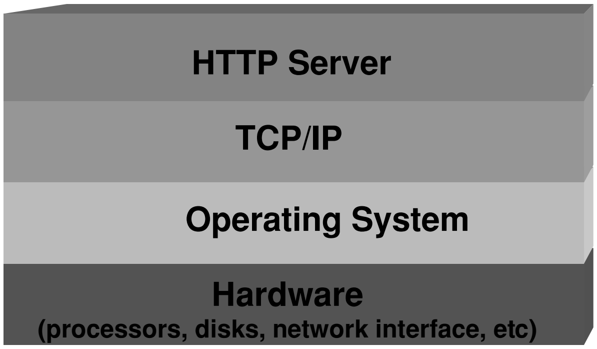


Figure 3: Layered infrastructure for Web-based systems.

UNDERSTANDING THE SYSTEM AND MEASURING PERFORMANCE

A complex infrastructure supports the delivery of Web services. The main components of any Web-based architecture are the hardware (central processing units [CPUs], disks, network interface cards, etc.), the operating system, the communication protocol (TCP/IP, UDP [uniform datagram protocol], etc.), and the HTTP server (Figure 3).

The bottleneck in performance can be anywhere along the chain of delivery through this infrastructure. It is therefore important to understand what to measure and how to measure to model and predict future needs accurately.

Web services infrastructure consists of not only servers that crunch user requests, but also local area networks (LANs), wireless area networks (WANs), load balancers, routers, storage systems, and security systems. To make matters worse, many of these resources are shared between many other Web services, groups, companies, and individuals.

Determining Response Time

The overall response time can then be broken down into two major components: the network time, which is the time spent traveling through the network, and the Web site time, which is the time the request spends being processed at the Web site. The network time consists of the transmission time of the request. This is the time it takes for the required information to be sent from the browser of the user to the Web site. This can vary depending on the technology the user has (modem, digital subscriber line [DSL], cable, etc.) as well as how much data is being sent (which dictates how many packets of information will be sent). Even the TCP/IP stack on the user side has an impact on the overall performance. There are many commercial and customer implementations of TCP/IP stacks. Benchmarking must be performed to get the true performance. The other component of network time is the latency, which is a measure of how many round-trip messages are required to be sent from the user to the Web site. Again, this varies and is dependent on many factors. For example, if the Web site requires the use of a cookie to be placed on the user machine, then each time the Web site is visited, the cookie exchange will take additional overhead to complete and must be added to the overall response time.

The main components of the Web site time include the service time and the queuing time. Each of these components is dependent on the hardware available to handle these functions—the CPU, the storage disks, and the LAN network. The service time is the time spent getting service from the CPU, storage device, and so on. Modern programming models protect these resources using mechanisms such as semaphores to prevent problems such as the shared data problem in which more than one task can corrupt data structures based on their calling sequence. Mechanisms such as semaphores inherently imply that while one task is using a resource like a CPU, other tasks that also want to use that resource must wait until the resource is free. This leads to a queuing model, which must be considered and modeled because it can add a substantial amount to the overall Web site time. Engineers must understand what the specific queuing model is (operating system, ping-pong buffer, etc.) and consider the impact in the overall performance numbers. In some network devices such as routers, the queuing mechanism may simply be a hardware buffer where all requests get stored and processed.

Configuration of network resources has a significant impact on the overall performance estimate. For example, adding fast memory (random access memory, or RAM) to a server will improve performance by some amount (which must be measured). Access time to RAM is much better (by orders of magnitude in some cases) than access time to magnetic media like a hard disk. The trade-off is usually cost and the cost-versus-performance analysis must be a variable that is known in advance (usually described in the SLA). Even the configuration of the user browser can have an impact on overall performance. Cache size settings in the browser, for example, can have an impact. Because this is a parameter set by the user, the model should consider either default values in the browser or be set to some average industry setting.

Web page download time must also be considered because this varies considerably based on the application. Modern Web pages contain markup language as well as embedded images and other embedded objects. To estimate the average download time for a Web page (independent of network traffic and other factors not associated with the Web page itself, although these factors must be considered in the final analysis), analysis must be performed on various computer configurations and settings. Models must be developed that take into consideration the number of embedded objects, the size of the embedded objects, HTTP header size, and the number segments per object. The more elaborate the Web page, the more processing is required to get the page to the user and provide the response time called for in the SLA. Keep in mind that the response time for a Web page with complicated embedded objects will vary considerably depending on the connection type. If many users will be working or accessing the site from home, dial-up modems' performance should be considered, not only the performance assuming a T1 line or other high-performance connection.

Interaction Performance

As with most user interface systems, it is not just the performance of the system itself that must be considered, but also the performance of the person interacting with the Web site. When analyzing any user interface

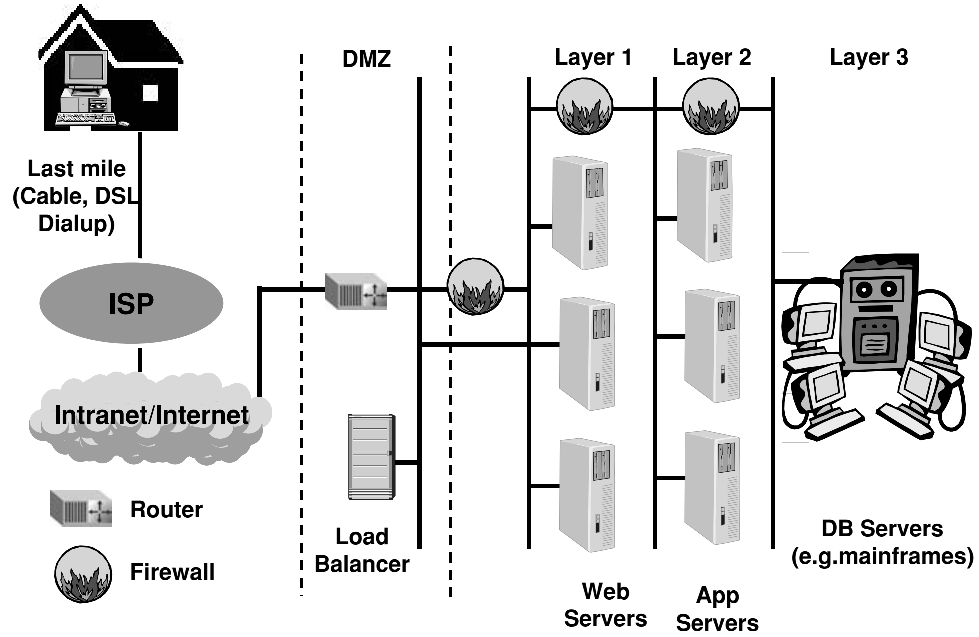


Figure 4: The Web infrastructure consists of many components, all of which must be considered for accurate capacity planning.

for performance-related issues, there is a delay associated with the interaction associated with the user think time. This is the period of time in which the user is perceiving the information and deciding what to so next. If the Web page is complicated and difficult to navigate, this "think" time will increase. Easy to understand and navigate Web sites reduce the think time. Designers of Web sites need to be careful when using the latest Web development technology to create fancy animated images, using multiple colors, and so on because these cause the user to become distracted from the task and lead to unnecessary delays in the overall interaction.

A Web infrastructure contains many interacting components (Figure 4). Servers, Internet service providers, firewalls, several levels of servers, load balancers, and so on combine in different ways to achieve a certain performance level. There can be a significant performance difference depending on whether the user is accessing information from inside or outside a firewall, for example. With the growing popularity of wireless technology and the Wireless Application Protocol (WAP), the complexity will continue to grow. This presents significant challenges with respect to performance analysis and capacity planning. The randomness with which the thousands or millions if users interact with a company intranet or extranet or the Internet, in general, makes the forecasting and capacity planning job extremely difficult. The first step is to understand the various components involved in the deployment of a system and model the current and predicted workloads as accurately as possible.

THE PERFORMANCE AND CAPACITY PROCESS

Menasce and Almeida (2002, p. 178) defined adequate Web capacity as having been achieved "if the Service Level Agreements (SLAs) are continuously met for a specified technology and standards, and if the services are provided within cost constraints." This assumes that the organization has an SLA. If not, adequate capacity will normally be defined by users who complain or stop using the service if they do not consider the performance adequate.

Model for Capacity Planning

There are many models for capacity planning. Figure 5 is one simple model that describes the major factors involved in capacity planning. Regardless of the model, before improvements can be made and plans for the future can be drawn up, there needs to be a way to assess current

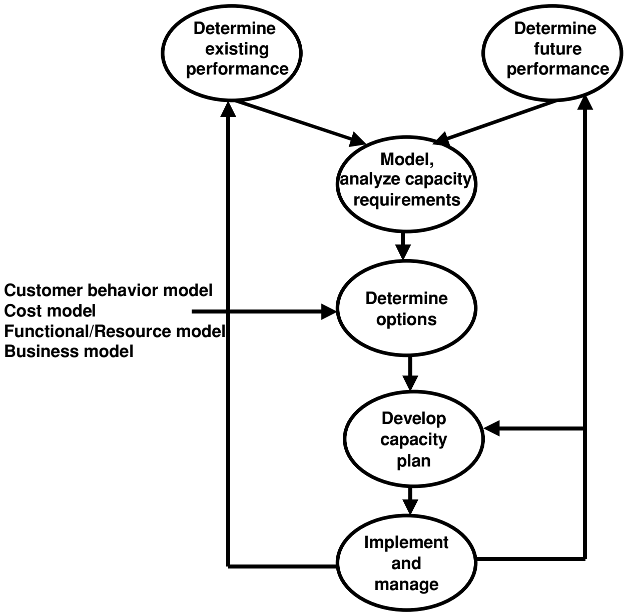

Determine existing performance
Determine future performance
Model, analyze capacity requirements
Customer behavior model
Cost model
Functional/Resource model
Business model
Determine options
Develop capacity plan
Implement and manage

Figure 5: The performance and capacity planning process.

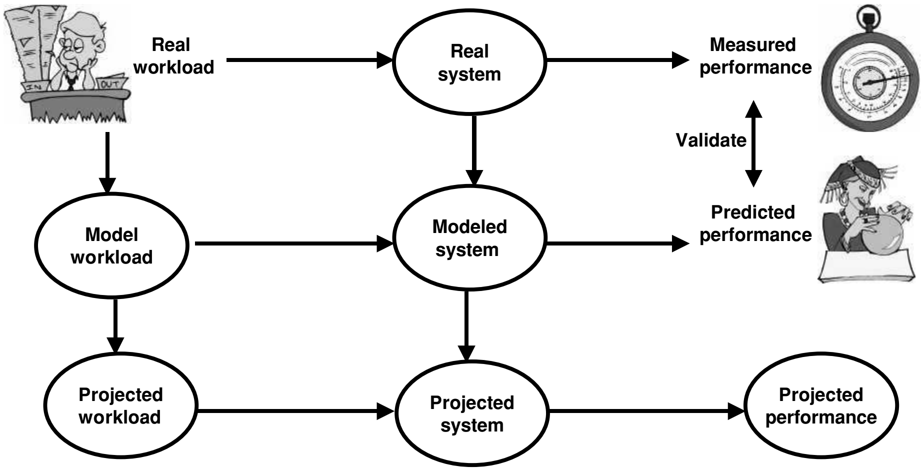


Figure 6: Modeling a Web system produces a platform by which to estimate future performance

performance. This requires understanding the system and environment and estimating its use. To understand the environment adequately requires an understanding of the client and server technology, the applications running on the servers and clients, the network connectivity used, the access protocols in the system, and the usage patterns of various types of customers and clients (e.g., when do they shop, when do they log on, what pages are frequented most often, etc.).

Usage Patterns

Once the main components of the system are understood, the usage patterns should be mapped onto the system model to determine the overall workload characterization. This provides the capacity planner an estimate as the workload intensity allocated to each of the main components in the system. This is useful for determining bottlenecks and knowing where to focus the effort in terms of overall system performance. Each of the components of the system can be measured in terms of the following parameters:

- Number of client visits,
- Average message size,
- Total CPU use,
- Number of transactions per unit time,
- Number of e-mails sent and received per day, and
- Number of downloads requested.

The actual parameters depend on the component being measured, and other components in addition to these can come into play as well. They are different depending on whether the component is an e-mail system, a search engine, an interactive training module, and son on. The capacity planner must determine what makes sense to measure for each of these components.

Once the components have been selected and the important parameters have been chosen for each component, the capacity planner must collect the appropriate data and perform a series of benchmarks to measure the performance on the actual physical machine. The main benchmarking technique is to run a set of known programs or applications in the target system and measure the actual performance of the system in relation to the parameters chosen in the analysis phase. It is not important to run a truly representative application, as long as the workload is well defined on a given system so an accurate comparison can be made. For Web-based systems the major benchmarking measures are time and rate. Time is from the users point of view; how long it takes to perform a specific task is important to this user group. From a management perspective, the main measure is the rate that drives how many users can be processed per unit time, which relates to overall productivity, revenue, or cost, depending on the application.

The process of benchmarking and measuring true performance on a real system is an important step and one that must be completed before proceeding to the step of modeling and predicting future performance. As shown in Figure 6, real workload should first be run on real system and the performance measured. The next step is to model the workload so that an accurate projection can be made about the future. This modeled workload must be run on the "system of the future," which is unknown at this time. To perform the required "what if" analysis, the system must also be modeled. This allows the modeled workload to be run on a modeled system. Measurements are made to ensure that the modeled system is an accurate representation of the real system. These measurements are made directly between the real system and the modeled system (of the real system). Any differences in the modeled measurements and the real measurements must be explained and understood. The modeled workload should be run through the modeled system to get the predicted performance. As a final validation step, the predicted performance produced by the modeled system should be compared with the measured performance of the actual system. Again, any differences should be understood.

Only after the model of the system has been validated can the effort of developing a projected workload and a projected system model be made. The projected workload should come from a variety of sources, including the SLA.

[Image not available in this electronic edition.]

Figure 7: Deciding how to model the Web system depends on the questions that are needed to be answered. From *Scaling for E-Business* (p. 289), by D. A. Menasce and V. Almeida, 2000, Upper Saddle River, NJ: Prentice Hall, 2000.

Modeling Parameters for Web-Based Systems

The key parameters associated with modeling a Web-based system are workload, performance and configuration. These parameters can be used in various ways to answer different questions. For example, analyzing workload plus configuration gives a measure of performance, which helps in predicting the future system requirements. Likewise, configuration and performance give a measure of workload, which helps to determine saturation points of the system. Finally, performance and workload help determine system configuration, which helps determine the sizing required in the future system.

Modeling Approaches for Capacity Management

Capacity planning can be simplified substantially by creating representative models of the real world using a simplified capacity model. This model should be based on critical or bottleneck resource availability as well as by interpreting the demand on that resource alone to determine the overall likely output. This technique can provide a rough order of magnitude verification that demand and capacity are in balance.

The phrase "all models are wrong but some are useful" is accurate in the sense that models, by definition, are abstractions of reality and therefore not 100% reflective and accurate with respect to reality. As the level of abstraction increases, the accuracy of the model decreases. System level models are higher levels of abstraction but can nevertheless be useful for modeling complex systems like Web-based systems. They can provide meaningful information to the capacity planner when making decisions about how to build and deploy future systems. Depending on the data needed for analysis, the question that needs to be answered is in what detail does the system (exiting and proposed) need to be modeled (Figure 7).

Of the many modeling approaches available, I focus here on several proposed by Menasce and Almeida (chapter 4), which are applicable to Web-based systems. The client–server interaction model is one approach that helps one to understand the interaction between the clients and servers in the system. In a multitiered system, this model can be useful in showing the important interactions at each tier. This can be used for future workload estimates. As shown in Figure 8, each e-business function of importance can be modeled in this way to show all possible interactions with that function. In Figure 8a, the interactions at the different computing tiers over time in a model that resembles a UML (unified modeling language) sequence diagram (where time increases from top to bottom) can be represented. This provides a time-based model of the interactions over time. Figure 8b shows the interactions between the different servers (application server, database server, and Web server) in a Markovian model. The Markovian model can be thought of as consisting of states (server nodes), arcs (arrows connecting the nodes showing interaction between the servers), and probabilities that represent the navigation patterns for the specific e-business function. This information effectively represents the customer behavior when interacting with the e-business function. Web sites can be optimized by applying modeling approaches such as the Markov model to the analysis of Web logs (Venkatachalam & Syed, n.d.).

Finally, the message sizes can also be represented and are shown in the diagram as well. With this information, a mathematical model can be developed that estimates the workload, message traffic, and other meaningful information about the system. Keep in mind that, like all other Markovian models, the model is only as accurate as the probability data assigned to the arcs. Relatively accurate

[Image not available in this electronic edition.]

Figure 8: A client…serverinteraction diagram showing all possible interactions for an e-business function: (a) the interactions over time; (b) the interactions showing navigation patterns and message sizes. From *Scaling for E-Business* (pp. 74…75)by D. A. Menasce and V. Almeida, 2000, Upper Saddle River, NJ: Prentice Hall, 2000.

data can be obtained from analyzing current systems, the logs on these systems for existing applications, estimates from other sources such as prototypes, and other products in the field. These models are also hierarchical, which allows them to be decomposed into lower level, more detailed models when necessary.

At the next lower level of detail (the component level), additional detail can be added. For example, queuing affects can be analyzed as service requests pass through the various levels of a multitiered application. Figure 7 can be represented from a queuing perspective as shown in Figure 9, in which each symbol represented a queuing function. Combination of these queuing functions form a queuing network in which each queue represents a system resource, such as a CPU or a disk drive, and the request waiting to use that resource. Average response times can be applied to each queue. Each of the system resources may have a different queue characteristic; use of the resource may be load independent which means the service time is not dependent on the queue length (e.g., a disk access), it may be load dependent where the service time is a function of the queue length (e.g., a CPU), or the queue may be a simple finite delay element (e.g., a network).

Regardless of the modeling approach used and which parts of the system are modeled, the result should be answers to the important questions raised during the planning process and used to drive the decisions on what and where to improve to meet future demands.

[Image not available in this electronic edition.]

Figure 9: A queuing model of the different service layers from Figure 8. From *Scaling for E-Business* (p. 289), by D. A. Menasce and V. Almeida, 2000, Upper Saddle River, NJ: Prentice Hall, 2000.

SOFTWARE PERFORMANCE ENGINEERING—MANAGING THE PROCESS

Many Web-based systems must meet a set of performance objectives. In general, performance is an indicator of how well a software-intensive system or component meets a set of requirements for timeliness. Timeliness can be measured in terms of response time, the time required to respond to some request, and throughput, which is an indicator of the number of requests that can be processed by the system in some specified time interval. Scalability is another important dimension of an embedded real-time system. Scalability is a measure of the systemís ability to continue to meet response time or throughput requirements as the demand for the system increases.

Choosing the right server, network, software, and so on for the job means nothing without proper performance management through the development life cycle. The consequences of performance failures can be significant, from damaged customer relations, to lost income, to overall project failure and even loss of life. Therefore, it is important to address performance issues throughout the life cycle. Managing performance can be done reactively or proactively. The reactive approach addresses performance issues by using a bigger server, dealing with performance only after the system has been architected, designed, and implemented and waiting until there is actually something to measure before addressing the problems. Proactive approaches to managing performance include tracking and communicating performance issues throughout the life cycle, developing a process for identifying performance jeopardy, and training team members in performance processes.

Definition of Software Performance Engineering

Software performanceengineering (SPE) is a proactive approach to managing performance. SPE is a systematic, quantitative approach to constructing software intensive systems that meet performance objectives (Smith & Williams, 2002). SPE is an engineering approach to performance, which avoids the “fix it later” mentality in designing real-time systems. The essence of SPE is using models to predict and evaluate system tradeoffs in software functions, size of hardware resources, and other resource requirements.

The Software Performance Engineering Process

The SPE process consists of the following steps (Smith & Williams, 2002):

- **Assess performance risk.** What is the performance risk of the project? The answer to this question helps determine the amount of effort to put into the SPE process.
- **Identify critical "use cases."** Use cases are an informal, user-friendly approach for gathering functional requirements for a system (Booch, Rumbaugh, & Jacobsen, 1999). Critical use cases are those use cases that are important from a responsiveness point of view to the user.
- **Select key performance scenarios.** These represent those functions that will be executed frequently or are critical to the overall performance of the system.
- **Establish performance objectives.** In this step, the system performance objectives and workload estimates are developed for the critical use cases.
- **Construct performance models**. A relevant model is developed for the system to measure performance. This can be an execution graph, a rate monotonic resource model (Rate Monotonic Analysis, 1997) or other relevant model to measure performance.
- **Determine software resource requirements.** This step captures the computational needs from a software perspective (e.g., number of messages processed or sent).
- **Add computer resource requirements.** This step maps the software resource requirements onto the amount of service required from key system devices in the execution environment (e.g., a server processor, fast memory, hard drive, router).
- **Evaluate the models.** If there is a problem a decision must be made to modify the product concept or revise the performance objectives.
- **Verify and validate the models.** Periodically take steps to make sure the models accurately reflect what the system is really doing.

SPE Assessment Requirements

The information generally required for a SPE assessment for network systems is as follows:

- **Workload—** the expected use of the system and applicable performance scenarios. It is important to choose performance scenarios that provide the system with the worst case data rates. These worst case scenarios can be developed by interfacing with the users and system engineers.
- **Performance objectives.** This represents the quantitative criteria for evaluating performance. Examples include server CPU utilization, memory utilization, and I/O bandwidth. The choice, in part, depends on the customer requirements.
- **Software characteristics.** This describes the processing steps for each of the performance scenarios and the order of the processing steps. One must have accurate software characteristics for this to be meaningful. This data can come from various sources such as early prototype systems using similar algorithm streams. Algorithms description documents, if available, also detail the algorithmic requirements for each of the functions in the system. From this, a discrete event simulation can be developed to model the execution of the algorithms.
- **Execution environment.** This describes the platform on which the proposed system will execute. An accurate representation of the hardware platform can come from a simulator that models the I/O peripherals of the embedded device as well as some of the core features. The other hardware components can be simulated as necessary.
- **Resource requirements.** This provides an estimate of the amount of service required for the key components of the system. Key components can include CPU, memory, and I/O bandwidth for each of the software functions.
- **Processing overhead.** This allows the mapping of software resources onto hardware or other device resources. The processing overhead is usually obtained by benchmarking typical functions (search engine, order processing, etc.) for each of the main performance scenarios. One example of a flow used to develop this data is shown in Figure 10.

The model is only as accurate as the data used to develop the model. For example, key factors that influence the processor throughput metric are as follows:

- The quantity of algorithms to implement
- Elemental operation costs (measured in processor cycles)
- Sustained throughput to peak throughput efficiency
- Processor family speed-up

The quantity of algorithms to perform is derived from a straightforward measurement of the number of mathematical operations required by the functions in the algorithm stream. The number of data points to be processed is also included in this measurement. The elemental operation costs measures the number of processor cycles required to perform typical functions. The sustained throughput to peak throughput efficiency factor derates the "marketing" processor throughput number to

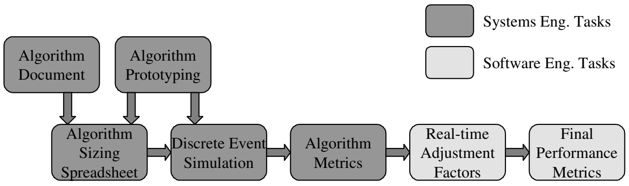

Systems Eng. Tasks
Software Eng. Tasks
Algorithm Document
Algorithm Prototyping
Algorithm Sizing Spreadsheet
Discrete Event Simulation
Algorithm Metrics
Real-time Adjustment Factors
Final Performance Metrics

Figure 10: Performance metric calculation flow.

something achievable over the sustained period of time a real world code stream requires. This factor allows for processor stalls and resource conflicts encountered in operation. The processor family speed-up factor can be used to adjust data gained from benchmarking on a previous generation processor.

Key factors that influence the memory utilization metric are as follows:

- Size and quantity of intermediate data products to be stored,
- Dynamic nature of memory usage,
- Bytes/data product,
- Bytes/instruction, and
- Size and quantity of input and output buffers based on worst case system scenarios (workloads).

SPE for Web-Based Applications

For Web-based applications, the SPE calls for using deliberately simple models of software processing. These should be easy to build and solve (like the ones discussed in the previous section) and provide the right feedback on whether the proposed software architecture will be sufficient to meet the performance goals. The SPE approach relies on execution models which then get mapped to system resources. The goal is to provide enough data to make the capacity planner comfortable with the following (Smith & Williams, 2002, p. 132):

- Placement of objects on the right processor and processes,
- Understand the frequency of communication among the objects,
- Understand which forms of synchronization primitives are required for each communication between the software objects,
- The amount of data passed during each communication, and
- The amount of processing performed by each software object.

AVAILABILITY MODELING AND PLANNING

When online systems are down, productivity and revenue is lost. The magnitude of the loss varies, depending on the type of system and how it is used. But the numbers can be in the hundreds of thousands per hour and even per minute. Online systems operating 24 hours a day, seven days a week, 365 days per year for international business opportunities have become a key mechanism for delivering services and products to customers. Downtime is just as important as a store being closed in the middle of a business week. Customers unable to access online systems are likely to take their business elsewhere, resulting in long-term revenue loss as well.

Although all downtime (also referred to as outage) is a potential loss of productivity or revenue (or both), some downtime is unavoidable. There must be "planned" downtime for system and application upgrades, new hardware and software, and backups. It is the "unplanned" downtime which must be minimized. Unplanned downtime occurs because of hardware failures, software crashes, computer viruses, and hacker attacks.

The solution to online system availability is not adding more hardware and other system resources. The system platform accounts for about 20% of the total system availability. The other 80% comes from a combination of people, process, and product (McDougall, 1999, p. 2).

Process

The industry has developed many processes over the last couple of decades to increase overall system availability. Some of the well-proven processes include system installation standards, change control, release upgrade processes, and backup and recovery testing. Just as standard software development processes are in place to allow quicker development of quality software, so are good processes and techniques important for maintaining online systems.

People

Availability should be considered an attitude instead of a priority. The staff responsible for maintaining the system should be trained properly to deal with backup and recovery techniques, as well as the standard processes for conducting business.

Product

The system platform itself contributes to overall downtime but not as much as the system process and people techniques described earlier. The system includes the hardware, the network infrastructure, and network operating system, and other required software, and hardware support. The investments made in the product (hardware and software) will add to the overall availability, but the right system configurations must be tested to ensure reliability of all the system parts working together. Given the large combination of different configurations, proper planning for this form of system testing is paramount to prevent unanticipated system surprises.

Availability Specification

Before beginning to address system availability, there must exist a set of requirements that define the key system goals for availability. An example availability specification may have the following statements:

- "During the peak hours of the system, 90% of queries will be completed in less than 1 second on average, with up to 100 users online. No queries will exceed three seconds" (modified from Cockroft & Walker, 1999, p. 31).
- "The order processing system for the XYZ online bookstore will be capable of sustaining 3,000 transactions per second during normal business hours."

Measuring Availability

As with any kind of process improvement effort, there must be a way of knowing whether the investment in process or product improvement is working. The cost of achieving additional availability can be extremely expensive, so knowing when to stop investing is important. Most measurements of availability are measured in percentage terms. Typically "five nines" availability, 99.999% is a often cited goal, but can be difficult to achieve.

Before beginning to address system availability, there must exist a set of requirements that define the key system attributes:

- Coverage—the normal business hours of the system. If the system is reporting stock exchange transactions, for example, the normal business hours will be a certain part of the business day and week. Availability should always be measured in terms of the coverage time only.
- Outage—This represents the number of minutes per month that the system can be down.

Availability is then measured as (coverage–outage)/ coverage × 100.

Availability can also be measured in terms of mean time between failures (MTBF—how long the system is up) and mean time to repair (MTTR—how long the system is down).

$$\text{Availability} = \text{MTBF}/\text{MTBF} + \text{MTTR}$$

As can be seen from this availability equation, as the MTTR approaches zero, the availability approaches 100%. As the MTBF gets larger, the MTTR begins to have lesser impact on availability.

Availability must be measured at several levels including the hardware platform level, the network level, and the application level—there can be failures in any one of these layers that can bring down the entire system.

Design Principles for System Availability

There are several general design principles that should be used when designing system to meet availability goals. These include but are not limited to the following (Zuberi & Mudarris, 2001):

- Select reliable hardware,
- Use mature and robust software,
- Invest in failure isolation,
- Test everything,
- Establish service level agreements,
- Maintain tight security,
- Eliminate single points of failure, and
- Availability modeling techniques.

There are approaches to modeling systems to produce availability estimates, but this is a difficult process because many system components that must be modeled are interrelated which makes accurate modeling more difficult. The common modeling approaches used to determine system availability include Markov models and Monte Carlo simulation techniques (Gilks, Richardson, & Spiegelhalter). These approaches model the system as a series of known system states and state transitions. Additional information such as time taken to go between system states and the probability of moving between system states are included in the model to improve accuracy.

TOOLS TO SUPPORT CAPACITY PLANNING

Capacity-planning software was created to help network executives and capacity planners plan for long-term growth and implement new initiatives. These same planning tools can also be used to help companies make their existing assets stretch further. A common thought with capacity planning is "I want to buy more," but this is actually a flawed understanding. Capacity planning is oftentimes not about buying more.

It is no surprise that as the importance of capacity planning has grown, commercial tools have become available to help manage the process, collect data, and visualize it in useful ways. When deciding on a capacity planning tool, it is important to understand how the tool will be used in the capacity planning process. Selecting the right tool for the job remains just as important as it always has.

Examples of Capacity Planning Tools

Sun Microsystems Resource Management Suite focuses on the capacity planning for storage systems. Some of the common storage-related problems are as follows:

- The inability for companies to keep pace with increasing storage demands;
- Storage is too expensive and complex to manage effectively;
- The inability to accurately plan, budget, and justify future storage needs; and
- The lack of data to support new storage architectures.

When planning for storage capacity, the common questions to answer are as follows:

- How much storage do I have today?
- How can I prevent storage related crashes?
- How can I predict future capacity needs accurately?

The tool provides the necessary infrastructure to forecast storage growth, manage heterogeneous storage, create and enforce storage policies, track usage, and create plans for future upgrades. Trend graphs are available to characterize usage patterns over selected time intervals.

Compaq's Enterprise Capacity and Performance Planner is a modeling tool used to predict the performance of both stand-alone and clustered systems. The tool is used to determine system performance levels for various workloads and system configurations. The tool also collects and analyzes data collected on the various platforms. Performance predictions are made with the tool using analytic queuing network models. A graphical component allows for "what if" analysis. A baseline model is developed from the data collected from the existing system and becomes the starting point for assessing the impact of changes to the system configuration of user estimated workloads.

The "what if" analysis forms the foundation for the overall capacity planning effort. Performance statistics are provided in detailed reports that encompass the main capacity planning performance statistics:

- Resource utilization,
- Response time,
- Throughput,
- Queue length, and
- Service time.

CAPACITY MANAGEMENT—PERFORMANCE AND STRESS TESTING OF WEB-BASED SYSTEMS

The only true way to assess the performance of a product is to try it out. The same holds true for a Web-based system. Performance testing is done to gain an understanding of the specific services being offered under specific and actual workload conditions. The capacity planning team should work in cooperation with other groups—development, production, and management staff—to plan, execute, and follow up on the results of this process.

Types of Performance Testing

There are several types of performance testing, including load testing which measures the performance of the system under load conditions called out in the SLA. The load can be actual or simulated depending on the system and process used. Spike testing is another useful performance test that tests a Web service under specific circumstances, and, just like the name implies, subjects the system to a heavy spike of traffic to determine how the system will react under this specific load condition.

Probably the most important type of performance testing is stress testing. Stress testing attempts to expose and help address the following important concerns:

- Stability issues—unexpected downtime and poorly written Web objects,
- Performance problems—locate bottlenecks and whether the application will handle peak loads, and
- Capacity planning—how many machines are needed to support usage.

Stress testing, if performed correctly, can help find and fix problems to avoid financial losses as well as ensure customer satisfaction. Stress testing is effective at locating the following bottlenecks:

- Memory,
- Processor,
- Network,
- Hard disk, and
- COM (Common Object Model) component.

Stress Testing Model for Web-Based Systems

A traditional stress test model is shown in Figure 11. The Web server is the system under test and is connected to a number of stress clients that simulate the workload by performing certain tasks, requests, operations, and so on from the Web server. The test is controlled by a controller stress client that directs the other stress clients as to the workload patterns while the test runs as well as collects the information needed to analyze the results of the testing process.

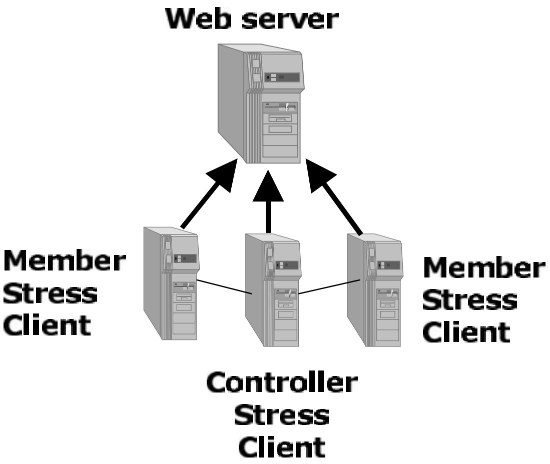


Figure 11: A traditional stress test model for Web-based systems.

The basic approach to stress testing involves the following steps:

- Confirm that the application functions under load,
- Find the maximum requests per second any application can handle,
- Determine the maximum number of concurrent connections the application can handle, and
- Test the application with a predefined number of unique users.

Based on the information obtained from the testing process, certain types of performance can be calculated. For example the following formula can be used to measure performance to aid in future capacity planning:

$$\text{MHz Cost} = N * S * \text{avg (PT)/avg(Rps)},$$

where N = number of processors, S = speed of processors, PT = % total processor time (this is a measure from the actual system servers), and Rps = requests per second, reports view (this comes from an analysis of the results of a test and is an application service provides (ASP) based measurement).

As an example, consider a test using a four processor Web server that achieved 750 requests per second, with the processors 80% utilized. This works out to

$$4 \text{ processors} * 500 \text{ MHz} \Longrightarrow 2 \text{ GHz}$$
$$80\% \text{ processor utilization} \Longrightarrow (2 \text{ Gig}) * (0.80) \Longrightarrow 1.6 \text{ GHz used}$$
$$750 \text{ ASP (Active Server Page) requests per second}$$
$$1.6/750 \Longrightarrow 2.1 \text{ million cycles per ASP request}$$

There are commercial tools available to aid the capacity planner in performing different types of performance testing, including stress testing. One such tool from Microsoft called Application Center Test (ACT) automates the process of executing and analyzing stress tests for Web-based systems.

The explosive growth of Internet sites and e-commerce has presented new challenges in managing performance and capacity. Also, the ability to estimate the future performance of a large and complex distributed software system at design time can significantly reduce overall software cost and risk. Capacity planning is the disciplined approach of managing system resources and defines the process of predicting when future load levels will saturate the system and of determining the most cost-effective way of delaying system saturation as much as possible.

GLOSSARY

Capacity planning The process of predicting when future load levels of an Internet-based system will saturate the system and the steps required to determine a cost-effective way of delaying system saturation as much as possible.

Client–Server A communication model between computer systems in which one system or program acts as the client and makes requests for service from another system or program called the server that fulfills the request.

E-commerce The buying and selling of goods and services on the Internet.

Markov Model A model consisting of a finite number of states and a probability distribution relating to those states, which governs the transitions among those states.

Network time The time spent for a packet of information to travel through a network.

Performance management An integral part of the capacity planning process in which system performance is managed proactively to ensure optimum efficiency of all computer system components so that users receive adequate response time.

Service level agreement An agreement between the customer and the system developer that outline the acceptable levels of performance and capacity requirements for the system.

Service time The time spent getting service from the central processing unit, storage device, and so on.

Software performance engineering A proactive, systematic and quantitative approach to constructing software intensive systems that meet performance objectives.

Stress testing The attempt to create a testing environment that is more demanding of the application than it would experience under normal operating conditions.

Unified modeling language (UML) A general purpose notational language used to specify and visualize software-based systems.

Web site time The time that a request for information spends being processed at a Web site.

CROSS REFERENCES

See *Client/Server Computing; E-business ROI Simulations; Electronic Commerce and Electronic Business; Return on Investment Analysis for E-business Projects; Risk Management in Internet-Based Software Projects; Web Services.*

REFERENCES

Booch, G., Rumbaugh, J., & Jacobsen, I. (1999). *The unified modeling language user guide*. Reading, MA: Addison-Wesley.

Cockcroft, A., & Walker, B. (2001). *Capacity planning for Internet services*. Santa Clara, CA: Sun Microsystems Press.

Gilks, W. R., Richardson, S., & Spiegelhalter. D. J. (Eds.). (1995). *Markov chain Monte Carlo in practice*. Boca Raton, FL: CRC Press.

McDougall, R. (1999, October). *Availability—what it means, why it's important, and how to improve it.* Sun BluePrints Online. Retrieved August, 2002, from http://www.sun.com/solutions/blueprints/1099/availability.pdf

Menasce, D. A., & Almeida, V. (2000). *Scaling for e-business: Technologies, models, performance, and capacity planning*. Upper Saddle River, NJ: Prentice Hall.

Menasce, D. A., & Almeida, V. (2002). *Capacity planning for Web services: Metrics, models, and methods*. Upper Saddle River, NJ: Prentice Hall.

Rate monotonic analysis keeps real time systems on schedule. (1997, September). *Electronic Design News*, 65.

Smith, C. U., & Williams, L. G. (2002). *Performance solutions, a practical guide to creating responsive, scalable software*. Reading, MA: Addison-Wesley.

Venkatachalam, M., & Syed, I. (n.d.). *Web site optimization through Web log analysis*. Retrieved July 2002 from http://icl.pju.edu.cn/yujs/papers/pdf/HMMS.pdf

Zuberi, A., & Mudarris, A. (2001, March). *Emerging trends for high availability.* Presented at the Chicago Uniforum Chapter. Retrieved from http://www.uniforum.chi.il.us/slides/highavail

Cascading Style Sheets (CSS)

Fred Condo, *California State University, Chico*

INTRODUCTION

Structure and Presentation in HTML

Hypertext markup language (HTML) provides for embedding structural information into text files. User agents (Web browsers) contain a default style sheet that controls the display of HTML. In the mid-1990s, the makers of Web browsers, principally Netscape and Microsoft, introduced presentational extensions into HTML, without regard to the wisdom of such extensions. These extensions gave Web page producers the ability to override the default browser styles. Examples of these extensions include the *font* element for specifying type styles and colors, and the *center* element for centering text and graphics horizontally. The result of using these presentational extensions is a commingling of presentation information with the structural markup and content of Web pages, or, worse, a substitution of presentational markup for structural markup.

Such a result may not be intuitively disadvantageous, but there are several problems that arise from the practice of mixing content with presentational directives. For example, there is no way to express the idea that all level-one headings (h1) throughout a Web site should be centered, set in a sans serif font, in red. Instead, the same directives must be specified along with every instance of an h1 tag. The result is that every page in the site carries redundant presentational information, which the server must transmit over the network each time. Maintaining the site becomes error prone and inefficient. If the designer decides to change the presentation of h1 elements to a serif typeface, for example, then someone must edit every page of the Web site. A person editing many pages is likely to miss a few pages or a few instances of the h1 tag; hence, the appearance and consistency of the site will degrade. Because a real site will have many more style rules than just this one example, the complexity and risk of error in maintenance are compounded. In the worst case, a naïve Web page producer may omit the h1 tags entirely. Such a practice makes it impossible for user agents to detect headlines, which has its most serious impact on vision-impaired users, who depend on user agents to list headlines by voice.

Consider an HTML page that uses *font* and *b* tags for headings: `<font size="4" face="Arial" color="#FF0000"><b>Text of heading</b></font>`. This code devotes more text to presentation than to content, and the text concerning presentation has to appear with every instance of a heading. With style sheets, the HTML code becomes much simpler: `<h1>Text of heading</h1>`. The HTML code is easier to understand, and it carries with it the important information that the text is a major heading. In the corresponding style sheet, this code appears (but only once for the entire Web site): `h1 {font-weight: bold; font-size: 125%; font-family: Arial, sans-serif; color:#F00;}`.

Cascading style sheets enable HTML authors to change the default specifications in a style sheet rather than in HTML code, so that new style rules may be applied to all instances of a specified HTML context throughout a site. In addition, cascading style sheets help preserve the structural context of HTML that indicates how pieces of content relate to one another. For instance, it is possible to grasp the structural difference between content within level-one heading tags and content within level-two heading tags. Level-one headers take precedence over level-two headers. It is not possible to grasp with certainty the structural differences between content wrapped in one set of presentation tags and content wrapped in another set of presentation tags by evaluating the HTML code. For instance, HTML authors will often forgo the list and list item elements in order to control the spacing of list items and the images used as bullets. This usually results in HTML code that cannot be recognized as describing a list. As well as helping authors maintain a separation of structure and presentation, CSS affords other benefits that were never available under HTML extensions.

Benefits of CSS

Separation of Presentational Information From Structure and Content

Much of the trouble associated with the embedded presentational directives of HTML is attributable to the commingling of content with presentational information. Because the only way to associate presentational HTML tags and attributes with their targets is to place them in close proximity in the same file, there is no way to disentangle them. CSS provides a mechanism whereby the association of content and presentation can exist without physical proximity. Designers can, as a result, physically and logically separate the content from its associated presentational styles. The most important benefits of CSS arise from this partition of content and style.

Centralization of Presentational Information

Because Web page producers no longer need to reduplicate style information for every page, they can collect general style rules into a single file or small group of files. HTML 4.01 and XHTML 1.0 provide mechanisms for associating a page with one or more style sheets. Once that link exists, it is no longer necessary to edit each Web page to effect style changes. Instead, the designer edits the central style sheet, and the change immediately takes effect throughout all pages that are linked to the style sheet. The process of changing a central style sheet is vastly simpler and less error prone than performing a complex set of edits on every page in a Web site.

Adaptability to Multiple Media

It is no longer safe to presume that users access the Web solely by means of a graphical Web browser such as Mozilla, Opera, or Internet Explorer. Web-access devices, like human users of the Web, are more diverse than ever. The Web needs to adapt to presentation via handheld computers, cell phones, computer-generated speech, and printers.

Designer Control Superior to Presentational HTML Extensions

Designers, too, need control of Web presentation. Their demands motivated early browser makers to pollute (as it turned out) HTML with presentational controls. For example, the *font* element enables Web producers to influence the face, size, and color of type. The drawback is that such presentation control is intermingled with the rest of the Web page such that changing the presentation requires tedious and error-prone editing. Moreover, presentation controls embedded in the page increase disk space requirements and network transmission times and obscure the structural relationships between the HTML elements of a page.

User Control of Display

Users of the Web, too, need control of presentation. This idea makes no sense in traditional print media, which are permanent, rigid, and fixed at the time of printing. Users need to adjust the Web's presentation to accommodate their special needs. For example, some users need to use large print or accommodate a deficiency in color perception. Such users benefit if they can override the designer's presentation styles; they would be excluded from the Web without such a capability.

Goes Beyond HTML

Throughout this chapter, reference is made to HTML elements and attributes, because HTML styling is the principal application for CSS. CSS, however, is not limited to HTML. With the appropriate software and hardware, CSS can style any markup language. For an example of a non-HTML style sheet, see the style sheet for RDF at W3C, http://www.w3.org/2000/08/w3c-synd/style.css.

CSS Standards

The World Wide Web Consortium has promulgated two recommendations for CSS. These are cascading style sheets level 1 (CSS1; Lie & Bos, 1999a), and cascading style sheets level 2 (CSS2; World Wide Web Consortium, 1998). CSS2 adds to CSS1 support for different media types, element positioning, and tables and for a richer set of selectors.

DOCUMENT VALIDATION

For a browser such as Mozilla or Internet Explorer to apply styles to the appropriate part of an HTML document, it must be able to analyze unambiguously both the HTML document and the style sheet. This requirement imposes a modest burden on authors of Web pages: Their pages must follow the actual rules of the HTML specification, and the style sheets must conform to the CSS specifications. Fortunately, the computer can test pages and style sheets for conformance to the standards, and it can even correct some HTML errors automatically. Using a code generator rather than coding "by hand" in a general text editor also helps prevent errors.

Both the World Wide Web Consortium (2001) and Quinn (2002) provide online HTML validators. The latter will recursively or batch validate up to 100 pages at a time. Both validators are Web front ends for Clark's (n.d.) *nsgmls* software, which runs on the Unix or Windows command line. The World Wide Web Consortium (2002a) provides a CSS validator for CSS levels 1 and 2. Finally, HTML Tidy (n.d.), software for many platforms, automatically cleans up common errors in HTML and can convert presentational HTML into embedded styles. Using such tools helps ensure that standards-conformant browsers will render pages as expected.

MEDIA TYPES

Media types apply styles according to the kind of display device on which a user is viewing (or hearing) your page. Media types are part of CSS2. There are four dimensions that characterize media types. Although many style capabilities work across media, some styles are specific to a particular medium. For example, there are no font styles for audio media, nor is there stereo separation for print, but both print and screen media support background colors.

Each of the discrete points on the four dimensions labels a media group. The four dimensions that characterize media types are as follows: continuous or paged; visual, aural, or tactile; grid or bitmap; interactive or static. Continuous media consist of uninterrupted streams, such as a scrolling window. Paged media have discrete segments, such as paper sheets for print, or stepped screens on a handheld system. Visual, aural, and tactile media are, respectively, seen, heard, and felt. Grid media consist of a grid or array of discrete character positions. Examples include traditional computer terminals and character printers. Bitmap media can freely represent graphics, letterforms, and glyphs, like the screen of a modern personal computer. Interactive media permit the user to interact with the display, as in traditional screen-based Web browsing. Static media do not allow interaction. Printed pages are an example of a static medium.

To associate a media type with a group of styles, you can create a separate style sheet file for each media type, or you can group them in a construct such as the following: `@screen {...}`. All the styles between the braces would apply to the screen medium. When creating separate style sheet files, specify the media type in the link or style tag in the HTML document with a media attribute, such as `media="screen"`.

There are nine media types in CSS2, plus a tenth type, *all,* which makes no distinction based on medium. Many style sheets comprise a base sheet associated with all media types, with media-specific style sheets that override styles or provide additional styles. The nine media types are, alphabetically, as follows: aural, braille, emboss, handheld, print, projection, screen, tty, and tv. CSS2.1 (World Wide Web Consortium, 2002b) will drop the aural media type and add speech and will split the aural media group into speech and audio.

In the sections below, each media type is briefly described and is characterized according to media group on the four dimensions mentioned above (using the CSS2.1 designations). If a dimension is not listed for a particular media type, the media type belongs to no group on that dimension. Some media types may occupy more than one group on a dimension, depending on the context or particular display device.

Speech. *Speech* styles apply to speech synthesis. Users of speech synthesizers include the visually impaired, those who must not divert their visual attention, such as the drivers of automobiles, and those who cannot use written words. In terms of the four dimensions, aural media are *continuous, speech,* and *interactive* or *static*.

Braille. *Braille* is for dynamic braille tactile devices, not for braille embossed in a fixed medium. Braille is *continuous, tactile, grid,* and *interactive* or *static*.

Emboss. *Emboss* is for braille printers. Emboss is *paged, tactile, grid,* and *static*.

Handheld. *Handheld* is for handheld devices. The CSS2 specification characterizes such devices as having small, monochrome screens and limited bandwidth. Already, the emergence of color handheld devices has overtaken the standard. No doubt some future revision of the standard will catch up with technological changes. Handheld is *continuous* or *paged, visual, audio,* or *speech, grid* or *bitmap,* and *interactive* or *static*.

Print. *Print* is for printing on traditional opaque media, most commonly paper. Print is *paged, visual, bitmap,* and *static*.

Projection. *Projection* is for paged presentations, whether on a computer projector or printed onto transparencies. It is not for the general case where a standard computer screen is enlarged via a projector (*screen* is the media type in that case). Projection is *paged, visual, bitmap,* and *interactive*.

Screen. *Screen* is for common computer screens. Screen is *continuous, visual* or *audio, bitmap,* and *interactive* or *static*.

Tty. *Tty* is for devices, such as teletypes, terminals, or some handheld displays, that use a fixed-width character grid. The pixel unit is not allowed for this media type, as such displays are not pixel addressable. Tty is *continuous, visual, grid,* and *interactive* or *static*.

Tv. *Tv* is for low-resolution displays with sound available, but with weak or clumsy scrolling controls. Tv is *continuous* or *paged, visual* or *audio, bitmap,* and *interactive* or *static*.

CSS BOX MODEL

Every element displayed on a visual medium has an associated box with various parts that influence the display. A diagram of the box model appears in Figure 1. The box model defines the following regions: content, padding, border, and margin. Only the content region is not optional. Each region defines a set of properties. Each region is described below, working from the center outward.

The *content* is the region where the element is displayed. Its dimensions correspond to the width and height properties of the element. This means that the box may be larger than the *content-width* and *content-height* dimensions. The overall height or width of the box is the sum of the height or width property and the adjacent padding, border, and margin.

The *padding* is the region surrounding the content. No foreground text or graphic appears in the padding, but background properties of the element, such as color, do appear in the padding region.

The *border* is the region where border properties are drawn. Some browsers may incorrectly continue the background into the background of the border, so beware. The specification says that the border's appearance shall depend solely on the element's border properties. Even though you might think of a border as a line of zero width, in the box model, it may have any arbitrary width.

The *margin* is a transparent region outside the border region. It separates the visible parts of an element's box

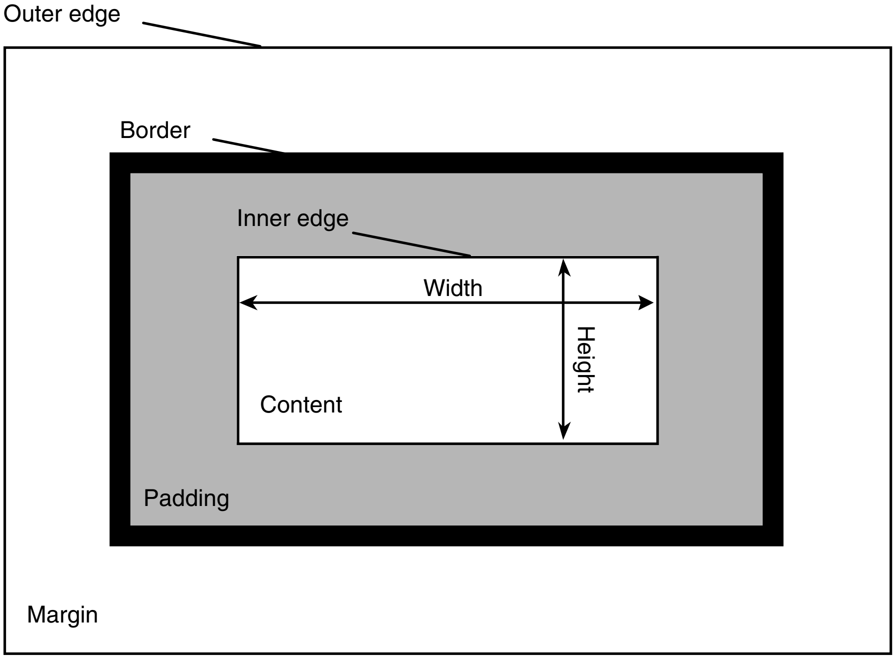


Figure 1: CSS box model.

from those of any adjoining boxes. Because the margin is transparent, the background of any enclosing element shows through.

Because HTML elements may contain other elements (as the *body* element contains a sequence of *p* elements), each box is nested within the box of its parent element. The entire box, including the padding, border, and margin, of the inner box, is inside the content region of the outer box. Thus, CSS boxes resemble nested Russian dolls.

Boxes do not overlap, with two exceptions. First, a box whose position has been altered through positioning properties may overlap other boxes on the page. Second, adjacent vertical margins, such as those of two paragraphs in sequence, collapse. You see only the larger of the two margins. Without the latter exception, the space between blocks of text would be excessive and tedious to control.

INHERITANCE AND THE CASCADE

In any given instance, several style rules may be "competing" to apply to a particular element of a Web page. This situation may seem chaotic, but allowing multiple style sheets to participate in styling Web pages is a major feature of CSS. It enables designers to create modular style sheets that are easy to maintain. It enables designers to override general styles for special cases. And it enables users to override designers so that the Web may accommodate special needs (for large type, for example).

A typical use of the cascade is to have two site-wide style sheets to which each HTML page has links in its *head* section. The first link refers to the style sheet for all media, and the second link refers to the style sheet for print. The print style sheet contains only *overrides* of the all-media style sheet. The remaining styles "flow" into the print style sheet (the term "cascading" is meant in analogy to a series of waterfalls). A page in the site that needs unique styles may have a *style* element in its head section. That style tag must appear *after* the *link* tags that refer to the overall style sheets for the site. By appearing last, the style rules in the style element can override the site-wide styles. Finally, a context requiring unique treatment may have a tag with a style *attribute*. The style attribute (not tag) is the most specific context of all and provides a local override of all other styles. Figure 2 shows a (contrived) HTML page that contains links to style sheets for multiple media, a style tag, and a style attribute.

When a browser assigns a value to a property for an element it is rendering, it goes through a process that, in principle, is quite simple. First, if the cascade (see below) yields a value, use that. Next, if the property in question is one that is inherited (see below), use the inherited value. Otherwise, use the initial, or default, value defined in the CSS specification.

Inheritance

Inheritance is based on an analogy to a family tree. Each HTML element that appears within some other element is said to be a *child* of the enclosing element. For example, in the HTML fragment of Figure 3, the emphasis element is a child of the paragraph element, and the paragraph is the child of the body. For properties that the specification says are inherited, descendents to any arbitrary level of descent (child, grandchild, great-grandchild . . .) inherit the property, unless the cascade overrides it. Elements that are children of the same parent element are called *siblings*.

Some properties are not inherited. This is for the convenience of the style sheet author. For example, it would be inconvenient in most cases if border properties were inherited, so they are not. Consider how tedious it would

```
<!DOCTYPE html PUBLIC  "-//W3C//DTD XHTML 1.0 Transitional//EN"
    "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml">
<head>
<link rel="stylesheet" type="text/css" href="all-media.css" media="all"/>
<link rel="stylesheet" type="text/css" href="print-overrides.css" media="print"/>
<style type="text/css">
    h1
    {
          font-family: "Times New Roman", Times, serif;
          color: yellow;
    }
</style>
<title>How Cascading Works</title>
</head>
<body bgcolor="#C0C0C0">
    <h1 style="font-family: Arial; color: blue">THIS HEADING1 TEXT
    IS BLUE ARIAL</h1>
</body>
</html>
```

Figure 2: Link, style-tag, and style-attribute contexts in the cascade.

be to have to turn off the inherited borders on each element within a paragraph with a border!

Because descent and inheritance are ambiguous when elements overlap, overlapping elements are not permitted in HTML. This is one of the reasons you need to validate your HTML code.

Cascade

Style Sheet Origin

Each style sheet has one of three origins: the style sheet's author, the user (who may have a personal style sheet), or the browser. Except when the *!important* weight is in play, the author's styles take precedence over the user's styles, which take precedence over the default styles of the browser. Browsers may not have !important rules.

Cascading Order

The cascade determines which style rule applies to a given situation by means of a four-step procedure.

1. Find all matching declarations from all style sheets involved. If there is but one matching declaration, use it and stop.
2. Otherwise, sort the matching declarations by origin and weight. Author styles override user styles, and user styles override browser styles. Any !important declaration overrides a normal declaration regardless of origin. If both author and user have an !important style, the user wins in CSS2 (the author won in CSS1, but this was an impediment to accessibility). If only one declaration is a clear winner, use it and stop.
3. Otherwise, sort by how specific the selector (see CSS Selectors, below) is. The more specific selectors override the less specific ones. If one style rule has the most specific selector, use it and stop.
4. Otherwise, if there is still a tie, the last rule wins. Apply it.

```
<body>
    <p>
      This is
      <em>very</em>
      important.
    </p>
    <p>
      I am a sibling of the
      paragraph above.
    </p>
</body>
```

Figure 3: Containment.

CSS SELECTORS

Every style rule consists of a *selector* and a set of *declarations,* each consisting of a *property* and a *value,* in this format: `selector { property: value; }`. There may be any number of declarations between the braces. The selector determines the context to which the declarations apply.

Selectors are patterns that match particular contexts in HTML documents. When a selector matches a context and makes it through the cascade, its style declarations apply to the whole context. For example, a selector "body" would match the context of the *body* element of an HTML document. For this reason, it is common practice to assign default values for the style sheet to the body selector. Through inheritance, those properties propagate to all the descendants of the body element, unless the cascading rules override them.

In addition to element names such as body, selectors may use the values of *id* or *class* attributes. Further, the specifications define a set of pseudoclass and pseudoelement selectors that behave like class and element selectors. They differ from real classes and elements in that they arise from a variable context or from user-generated events. For example, the *:first-line* pseudoelement depends on the width of the browser's display, and the *:visited* pseudoclass depends on the user's browsing history.

In addition, selectors may be grouped to apply a given style to several disjoint contexts. Finally, selectors may be arranged in various ways to match very specific contexts.

CSS1 defines a basic selector model that supports element, class, id, and descent contexts, including some pseudoelements and pseudoclasses. To this model CSS2 adds additional pseudoelements and pseudoclasses, as well as more precise contexts based on sibling and descent relationships between elements.

CSS1 Selectors

CSS1 defines the following kinds of selector: type selectors, descendant selectors, class selectors, ID selectors, link pseudoclasses, and typographical pseudoelements.

Type Selectors

Type selectors are so designated because they match all elements (tags) of a single type. A more intuitive way to regard such selectors is as redefiners of the appearance of HTML tags. Any HTML element, such as *body,* may act as a type selector.

```
body { background-color: #FFFFFF; }
```

Descendant Selectors

A space-separated list of selectors matches if the context named on the left contains the context named on the right. The sample code below sets emphasized text in red, but only if the emphasized text is in a level-two outline heading.

```
h2 em { color: #FF0000; }
```

Class Selectors

A class selector is introduced with a dot and matches the value of an HTML tag's class attribute. Example 1 below applies a border to any element whose class is "note"; example 2 does the same, but only if the element is a paragraph. Example 1:

```
.note { border-width: thin; border-style:
  solid; }
```

Example 2:

```
p.note { border-width: thin; border-style:
  solid; }
```

ID Selectors

ID selectors work the same way class selectors do, but they match the id attribute. In the style sheet, they are introduced with the # sign rather than a dot. Bear in mind that IDs must be unique in the HTML document; often a class is more convenient. Example 1:

```
#note1 { border-width: thin; border-style:
  solid; }
```

Example 2:

```
p#note1 { border-width: thin; border-style:
  solid; }
```

Link Pseudoclasses

The link pseudoclasses are *:visited* and *:link*. They correspond to links that the user has or has not followed. These pseudoclasses replace the link and vlink attributes of the HTML body element.

```
a:link { color: #0000FF; }
a:visited { color: #FF00FF; }
```

Typographical Pseudoelements

The typographical pseudoelements in CSS1 are *:first-letter* and *:first-line*. They apply styles to the corresponding parts of an element. In the case of :first-line, it dynamically adapts to changes in the window size or font. The example below creates a drop-cap on a paragraph designated with the id "initial".

```
p#initial:first-letter { float: left;
  font-size: 3em; }
```

CSS2 Selectors

CSS2 adds to CSS1 the following kinds of selectors: child selectors, the universal selector, adjacent-sibling selectors, dynamic pseudoclasses, the language pseudoclass, text-generating pseudoelements, and attribute selectors.

Child Selectors

Child selectors are pairs of element names separated by >, in which the right-hand element must be a child of the left-hand element. The first example sets the line height of paragraphs that are children of the body (child paragraphs of div would be unaffected). The second example sets ordered lists to have uppercase Roman numerals as markers, and ordered lists nested one level deep in other ordered lists to have uppercase alphabetic markers. Example 1:

```
body > p { line-height: 1.5; }
```

Example 2:

```
ol { list-style-type: upper-roman; }
ol > li > ol { list-style-type:
  upper-alpha; }
```

Universal Selector

The universal selector ∗ is a wild-card selector. It matches any HTML element name. The example below sets in red any emphasized text that is exactly a grandchild of the body element, regardless of what the parent of *em* is.

```
body > * > em { color: #FF0000; }
```

Adjacent-Sibling Selectors

Adjacent siblings are elements that have the same parent and no other elements between them. The general form of an adjacent-sibling selector is $x + y$, where x and y are elements. The style applies to y. The example applies a drop-cap to a paragraph if and only if it immediately follows a level-one heading.

```
h1 + p:first-letter { float: left;
  font-size: 3em; }
```

Dynamic Pseudoclasses

The dynamic pseudoclasses are *:hover, :active,* and *:focus*. Hovering occurs when the user points at but does not activate an element, as when pointing but not clicking on a link. An item is active while the user is activating it, such as during the time between clicking and releasing the mouse button on a link. Focus occurs when an element is able to receive input from the keyboard (or other text input device). In the example below, links are blue when unvisited, red when visited, yellow when pointed at, and green during the mouse click. Note that the order of the style rules is significant, because some of the selectors are equally specific. This cluster of style rules should appear in the order shown or some of the styles will never show up.

```
a:link    { color: blue; }
a:visited { color: red; }
a:hover   { color: yellow; }
a:active  { color: green; }
```

Language Pseudoclass

The language pseudoclass is *:lang(C),* where C is a language code as specified in the HTML 4 standard (Raggett, Le Hors, & Jacobs, 1999), or RFC 3066 (Alvestrand, 1995). (The CSS specification actually refers to the obsolete RFC 1766, but implementations should use the current definition of language tags.) Some instances of language codes are *en* for English, *en-us* for United States English, and *fr-ch* for Swiss French. The pseudoclass selector matches according to a liberal algorithm such that :lang(en) would match either of the English codes shown above, just as :lang(fr) would match the Swiss French code. The language of an element may be determined from any of three sources in HTML: a lang attribute on the element, a *meta* tag, or HTTP headers. Other markup languages may have other methods of specifying the human language of a document or element.

The example below is quoted from the CSS specification section 5.11.4 (Bos, Lie, Lilley, & Jacobs, 1998). It sets the appropriate quotation marks for a document that is either in German or French and that contains quotations in the other language. The quotation marks are set according to the language of the parent element, not the language of the quotation, which is the correct typographic convention.

```
HTML:lang(fr) { quotes: '« ' ' »' }
HTML:lang(de) { quotes: '»' '«' '\2039'
  '\203A' }
:lang(fr) > Q { quotes: '« ' '»' }
:lang(de) > Q { quotes: '»' '«' '\2039'
  '\203A' }
```

Text-Generating Pseudoelements

The *:before* and *:after* pseudoelements inject text before and after the content of the selected element. The sample code below inserts the bold label "Summary:" at the beginning of every paragraph whose class is "summary." Note the space between the colon and the closing quotation mark.

```
p.summary:before
{
      content: "Summary: ";
      font-weight: bold
}
```

Attribute Selectors

Attribute selectors match according to an element's attributes in one of four ways: *[attribute]* matches when the element has the specified attribute, regardless of value; *[attribute = value]* matches when the element's attribute has the exact value specified; *[attribute ~ = value]* matches when the attribute can be a space-separated list and one of the members of the actual list is the specified value; *[attribute | = value]* matches the value according to the rules for matching language codes, as specified in RFC 3066 (Alvestrand, 2001).

The example below applies a dotted border below any abbreviation or acronym element that has a *title* attribute. In addition, it asks the browser to change the mouse pointer to the help cursor when hovering over the element. Few browsers successfully render all these styles, however.

```
abbr[title], acronym[title]
{
    border-bottom: black;
    border-width: 0 0 1px 0;
    border-style: none none dotted none;
}
abbr[title]:hover, acronym[title]:hover
{
    cursor: help;
}
```

Grouping

To apply the same style to a set of distinct selectors, separate the selectors with commas. The example below applies a green text color to all six levels of heading.

```
h1, h2, h3, h4, h5, h6
{
      color: #006666;
      background-color: transparent;
}
```

CSS PROPERTIES

Every style rule comprises a selector and one or more declarations. Each declaration has a property and a value. The values permitted for each property depend on the domain of that property. For example, the *margin* property accepts only length values, and the *font-family* property accepts only a list of font names as its value.

Many properties have two forms, a specific form and a combined, or shorthand, form. For example, the *border* property sets the width, style, and color on all four sides, so it is much more compact to write than it is to write all the individual properties pertaining to borders. The shorthand form, however, has a subtle yet important impact on the cascading rules. Any properties that you do not set in the shorthand form act as though you had explicitly set the initial, or default, value in the style rule. Recall that although the cascading rules give low priority to default values, explicit rules (even implicitly explicit ones) take on the priority of the associated selector. You need to be aware of this subtle difference when you choose to use shorthand properties.

The CSS specifications define over 100 properties, and space does not permit them to be listed here. The property definitions are readily available online (Lie & Bos, 1999b; Bos et al., 1998) and in exhaustive references, such as Meyer (2000) and Lie & Bos (1999a). If you do refer to the specifications themselves, be sure to refer to the associated errata as well.

Value Types, Units, and Representations

Color

Color has four representations in style sheets: color names as defined in HTML 4; hexadecimal color codes (long and short forms); percentage RGB codes; numeric RGB codes. The following examples all represent the same color.

```
yellow
#FFFF00
#FF0
rgb(100%, 100, 0%)
rgb(255, 255, 0)
```

URL

Any URL (uniform resource locator) may appear in the following notation: `url()`. Between the parentheses, either absolute or relative URLs may appear. If document relative, they are relative to the style sheet itself.

Length

You may express absolute lengths in terms of inches (in), centimeters (cm), millimeters (mm), points (pt), or picas (pc). When writing a length, put no space between the number and the unit, as in this example: `3mm`.

You may express relative lengths in terms of ems (em), exes (ex), or pixels (px). The em unit is equal to the current font size. The ex unit is equal to the height of the lowercase letter *x* in the current font, although browsers often set it equal to 0.5 em. The pixel unit is usually a screen pixel, but for printing, the specification calls for the pixel unit to be rescaled so that it has the same approximate viewing size as it does on the screen.

Percentage

Percentages appear as a number followed by the percent sign, as in this example: `150%`. Percentage values for lengths are relative to some other length, usually some dimension of a parent element.

Key Words

All other values discussed in this chapter take key words or lists of key words as their values. For example, this style rule asks the browser to set the type of a page in a serif font, preferably Palatino or Times: `body { font-family: Palatino, Times, serif; }`.

PRACTICAL CSS IN ACTION

Linking Styles to Web Pages

There are two basic ways to associate a style sheet with a Web page. The first uses the HTML link element; the second uses the style element. Every page in a Web site needs to be associated with the site's style sheet, but once the association exists, maintaining the styles is easy.

Whether you use the link or the style element, it must be a child of the head element of your HTML page. The first example below uses link; the second uses style. Both examples are in HTML 4. Example 1:

```
<link rel="stylesheet" href="css/screen.css"
  type="text/css" media="screen">
```

Example 2:

```
<style type="text/css" media="all">
   @import url(css/screen.css);
</style>
```

W3C Core Styles

The W3C core styles are eight prewritten style sheets that are available for anyone to use with their HTML documents (Bos, 1999). They are called *Chocolate, Midnight, Modernist, Oldstyle, Steely, Swiss, Traditional,* and *Ultramarine*. You can interactively preview the eight style sheets at http://www.w3.org/StyleSheets/Core/preview. To use one of these style sheets, use a link element like the one in the example, which specifies the Modernist style sheet.

```
<link rel="stylesheet" href="http://www.
  w3.org/StyleSheets/Core/Modernist"
  type="text/css">
```

Alternate Style Sheets

Every Web page that uses style sheets has up to three different kinds of style sheet: persistent, preferred, and alternate, none of which is mandatory (Raggett, Le Hors, & Jacobs, 1999). The alternate style sheets are mutually exclusive, and modern browsers such as Mozilla afford the user a way to choose which alternate style sheet to use. The preferred style sheet is the one that the browser loads initially. The browser always combines the currently selected preferred or alternate style sheet with the persistent style sheet, if any. The persistent style sheet is a good place to put basic styles that the designer wishes to appear at all times. Then only the variant parts of the style sheet need to appear in the preferred and alternate style sheets. There may be any number of alternate style sheets but no more than one persistent style sheet and no more than one preferred style sheet.

The designation of a style sheet as persistent, preferred, or alternate depends on the particular combination of *rel* and *title* attributes on the link tag that associates a Web page with the style sheets: The link to a persistent style sheet has `rel="stylesheet"` and no title attribute at all. The link to the preferred style sheet has `rel="stylesheet"` and a title attribute of your choosing. The link to an alternate style sheet has `rel="alternate stylesheet"` and a distinct title of your choosing. Not all Web browsers provide a user interface for switching style sheets. To provide a switching facility to users of such browsers, you may use HTML form elements, such as buttons or pop-up menus, and some JavaScript code (Sowden, 2001; Ludwin, 2002).

Examples

The following examples are conservative. They are intended to show that interesting results can issue from quite simple CSS code. For more elaborate and cutting-edge examples, see the bibliography for items by Meyer (2001a; 2001b; 2002a; 2002b).

Logo Fixed to the Top Left Corner

The example in Figure 4 assumes that you have a 72-pixel-wide logo in a file *logo.png* that you wish to affix to the top, left corner of your page. The left margin is set in pixel units to move text to the right so that it does not overlap the logo. The margin is set somewhat larger than the width of the graphic so that there is some space between the logo and the text.

Two-Column Layout Without Tables

The examples in Figures 5 and 6 constitute a simplified version of the technique of Zeldman (2001). The goal here is to create a two-column layout with a list of links on the left and the page's principal content on the right. To achieve this layout, the HTML code in Figure 6 uses two div tags to organize the menu links and the content. The division with the content has an id of "content," to which the CSS code in Figure 5 refers. That code primarily employs the underutilized *float* property to float the content to the right side of the page. Margins and padding are adjusted for esthetics, and a border helps the eye distinguish the content from the menu links.

```
body
{
      margin-left: 82px;
      background-position: left top;
      background-attachment: fixed;
      background-repeat: no-repeat;
      background-image: url(logo.png);
      color: #000000;
      background-color: #ffffff;
}
```

Figure 4: Code for a fixed background image.

```
div#content
{
      float: right;
      border-left: 1px solid #000;
      border-bottom: 1px solid #000;
      width: 70%;
      padding-top: 0;
      padding-right: 0;
      padding-left: 3em;
      padding-bottom: 3em;
      margin: 0;
}
```

Figure 5: CSS for two-column layout.

BEST PRACTICES

Modern Web browsers are fairly good implementations of CSS (Web Standards Project, n.d.), and the CSS specifications offer a broad range of perhaps alluring stylistic capabilities. Nonetheless, adhering to certain best practices will minimize the risk that one or another of the standards-compliant browsers will fail to render a page as the designer expected. In addition, many of these practices result in fluid documents that adapt well to different window or screen sizes and user preferences for font size. Some of the practices avoid programming errors that exist in popular Web authoring agents. We examine best practices in two groups: (a) a list of practices by the inventors of CSS (Lie & Bos, 1999), and (b) some new practices that I recommend.

Recommendations of Lie & Bos (1999a)

Use Ems to Specify Lengths

Designs that use ems are scalable, fluid, and adaptive. The em unit, traditionally the width of the glyph M, in CSS is equal to whatever the height of the current font is. Because

```
<body>
<div id="content">
  <p>Content goes here.</p>
</div>
<div id="links">
  <ul>
    <li>link a</li>
    <li>link b</li>
    <li>link c</li>
    <li>link d</li>
  </ul>
</div>
</body>
```

Figure 6: HTML code fragment for two-column layout.

the font height is the basis of this unit, it is a relative unit that scales proportionally as the font size changes. When you set margins and padding in terms of ems, the margins and padding grow or shrink in proportion to the changes in the font on the display device. This preserves the balance and overall appearance of the design while accommodating a variety of display devices and user needs and preferences for font size.

Use Ems to Specify Font Sizes

This recommendation may seem oddly circular at first glance. What does it mean to specify font sizes in terms of font sizes? The em unit for font size is relative to the default font size of the Web browser or other user agent. For example, 2em is equal to twice the default font size. Using ems, rather than traditional type units such as points or picas, results in pages whose font size adapts to the needs and preferences of the end user, or to the limitations of the display medium. Using absolute units for font size is a common reason that type on Web pages appears too big or, worse, too small to read.

Use % When You Wish Not to Use Ems

There are elements for which em is not the appropriate unit. For example, the margins of the page as a whole (the body element) are usually more pleasing if they are based on the size of the display window, rather than on the font size. The % (percent) unit is the appropriate unit for such cases.

Use Absolute Units Only When You Know the Size of the Output Medium

First, a caveat: It is rare to know the dimensions of Web-related output media. Even in print, there are regional variations in the dimensions of the paper. The United States generally uses paper that is 8.5 by 11 or 17 inches for office productivity applications. In Europe, the A4 metric size is common. Bear in mind as well that even in print, some users may need larger than usual type, and others may prefer very small type. If, however, you do know the dimensions of the medium, and if your design has very tight tolerances, CSS provides a menagerie of absolute units: points, picas, centimeters, millimeters, inches.

Float Objects Rather Than Using Layout Tables

In presentational HTML, it is possible to float images or tables to the left or right margin; other content wraps around the floated elements. Designers often resort to a table to move other, nonfloatable elements to one side of the page, creating a multicolumn layout.

With CSS, any element can float if you assign the float property to it, so it is not necessary to use tables for layout. The two principal advantages of CSS floating over layout tables are faster rendering and better accessibility. Tables, especially complex layout tables, often take much time for the browser to render. In addition, the World Wide Web Consortium (1999) recommends against using tables for layout.

Arrange Content in Logical Order

Another way of stating this guideline is don't rely on CSS to position your content in the order you wish users to read it. Arranging the content of a page in its logical sequence ensures that the page continues to make sense even in browsers that cannot render the styles.

Test Pages in the Absence of CSS

Even in a modern CSS-compliant browser, the style sheet may become temporarily unavailable due to a network failure or other cause. For this reason, you should test your pages for legibility with style sheets off. Failure of pages to degrade gracefully in the absence of style sheets causes serious problems of accessibility as well.

Test Pages in Relevant Browsers

Despite the emergence of browsers that adhere to the HTML and CSS standards, it is still necessary to test designs in several relevant browsers. This is due to ambiguities in the standards, differences in interpretation of the standards, or outright errors in implementations. If a significant proportion of your site's visitors uses older browsers, testing is imperative.

Use a Generic Font Family

In print, designers can choose from a multitude of typefaces and can specify the font that best solves a communication problem. On the Web, only the typefaces installed on the end user's computer are available. Typically, these are the fonts installed by default in the computer's operating system. Nonetheless, you may wish to specify fonts for various contexts in your Web pages. When you do, always list the appropriate generic font family as the last item in the list of font families for your style rule. For example: `body { font-family: Palatino, "Times New Roman", serif }`.

There are five generic font families: *Serif* fonts have small extensions (serifs) on the ends of strokes; *sans-serif* fonts lack serifs (from the French sans, meaning "without"); *monospace* fonts have glyphs that are all the same width; *cursive* fonts look like handwriting; *fantasy* fonts are miscellaneous display fonts for special uses. Many computers lack cursive or fantasy fonts, in which case the browser will substitute another font.

Use Numeric Color Codes

The CSS standard defines only 16 color names (aqua, black, blue, fuchsia, gray, green, lime, maroon, navy, olive, purple, red, silver, teal, yellow, and white). Even though browsers may recognize other names, such names are nonstandard and may produce unexpected results, particularly in older browsers. Therefore, it is best always to specify colors as hexadecimal or percentage codes.

Know When to Stop

Listen to your graphic designer and your usability expert. CSS gives you broad power to control the formatting of Web pages. Such power calls for a trained eye and great restraint. You can create a page with 10 fonts and every word in a different color, but such a page would be unreadable. Usually two or three fonts suffice to distinguish the various kinds of text. Using color sparingly makes the color you do use stand out with great emphasis (but be sure that color is not the only way your page conveys information).

Other Best Practices

Use External Style Sheets

Create style sheets as discrete text files that you link to your Web pages. This practice eliminates unnecessary redundancy in the Web site and enables you to establish alternate style sheets. Use embedded styles only when you need to override your external style sheet for a single, exceptional page. If you find that you are often overriding the external style sheets, it is time to redesign the external style sheets. Finally, reserve inline styles with the style attribute only for situations where it is absolutely necessary to accomplish an effect, such as dynamically moving layers around on the page.

This practice is particularly important because it minimizes maintenance tasks. The less style information is dispersed throughout a site's pages, the less effort needs to go into maintaining the site's style sheets. By keeping as much style information as possible in centralized external style sheets, you will reduce the likelihood that a style you want to change is neglected.

Use @import to Hide Styles from Netscape 4

Version 4 of the Netscape browser is now badly obsolete, yet it still has many users. Its CSS implementation is both nonstandard and erroneous. It is possible to create valid HTML and CSS code that is completely unusable in Netscape 4, such that links become inactive. To avoid having to work around the problems in Netscape 4, eliminate styles for that browser by taking advantage of Netscape 4's ignorance of the *@import* directive.

Zeldman (2001) recommends code like the example below for screening style sheets from Netscape 4.

```
<style type="text/css" media="all">
    @import "/styles.css";
</style>
```

For XHTML, use code like the example below, which permits the page to validate under the XHTML document type definitions.

```
<style type="text/css" media="all">
    /*<![CDATA[*/
    @import url(/styles.css);
    /*]]>*/
</style>
```

You may need to decline this practice if your audience has a high proportion of Netscape 4 users. The trade-off is that you will be limited in what aspects of CSS you can use and you will suffer exposure to the risk that trivial changes in your CSS will break your site for Netscape 4. Rigorous testing is required under such circumstances. Because of the severity of the trade-offs, you may decide to politely encourage your users to move to browsers that support standards. The Web Standards Project (n.d.b) provides techniques for doing so.

Use a W3C Core Style

If you are not a graphic designer or do not wish to take the time to develop your own style sheets, use one of the W3C core styles mentioned above. They provide a variety of basic style sheets appropriate for general use, they have already been written and tested for you, and they work (as much as possible) in Netscape 4. If you use a core style, you may omit the preceding practice.

Judiciously Use and Name Classes

Use HTML classes to distinguish elements, such as paragraphs, that serve different purposes in your documents. Name the classes according to their function, not according to their appearance. You may change your mind as to the appearance, and it would make no sense for a class called "blue" to appear in green. But a class called "summary" will make sense regardless of the styles you apply to it.

Avoid Workarounds

One of the benefits of Web standards is that Web page creators can avoid working around quirks and bugs in browsers. Nonetheless, there are bugs and quirks even in modern browsers. Searching the Web for "CSS workaround" will confirm this. Many of the workarounds have to do with layouts that require pixel accuracy, and whose designers consider a browser's misinterpretation of the box model to "ruin" the design. To avoid such bugs, many examples on the Web use bizarre syntax that risks failure when repaired versions of browsers emerge.

Rather than work around the bugs, it is best not to insist on control down to the pixel. Heed the practices above that encourage fluid, adaptive documents, and heed the next practice, as well.

Be Flexible to Keep It Simple

CSS does not offer the precision of presentation control that exists in print. It would be counterproductive for the Web to have such a rigid level of control, because of the diversity of users who, and output devices that, access the Web. The best design practice is to create flexible presentation styles. Simplicity is best. It is not necessary to exercise every nuance of the box model or work around every bug in Internet Explorer. If you are willing to accept some deviation from your ideal design, your style sheets will be simple and easy to maintain.

EVOLUTION OF CSS

The World Wide Web Consortium (2002b) continues to develop the CSS standards. Most recently, it released a draft specification for CSS2.1. This update corrects several errors in the published standard for CSS2. More important, it amends the standard to conform with practice in the following ways: (a) maintains compatibility with widely implemented and accepted parts of CSS2; (b) modifies the specification where implementations have overwhelmingly contradicted CSS2; (c) omits features of CSS2 that have never (or almost never) been implemented; and (d) omits CSS2 features that are to be superseded in CSS3.

CSS3 is still under development. A key feature of CSS3 is that it is being developed in modules. This method of developing the latest CSS specifications will enable implementers and users to accept or reject parts of CSS3 without necessitating revisions to the entire specification. It would not be surprising if only small portions of CSS3 ever see widespread implementation. Nonetheless, it does address the needs of linguistic and other communities whose ways of communicating were ignored in CSS1 and CSS2.

GLOSSARY

Border The region of the box model immediately outside the padding.

Box Model A part of the CSS specification that defines where the content and other parts of an HTML element will appear on the display medium.

Cascade The rules or logic by which multiple style sheets participate in an orderly manner to affect the appearance of a document.

Child An HTML element that is part of the content of an enclosing HTML element.

Content The region of the box model where an HTML element's text or graphic appears.

Declaration The assignment of a value to a property.

Inheritance The propagation of a style from a parent to a child or other descendant.

Margin The transparent region of the box model surrounding the border.

Padding The region of the box model between the content and the border; background properties appear within it.

Parent An HTML element that contains another HTML element.

Property An aspect of style that CSS can control.

Selector The part of a style rule that expresses the context to which the style should apply.

Style Rule A complete expression in CSS that specifies styles and the context to which they should apply; consists of a selector and declarations.

User Agent Software or a device that acts on behalf of a user. The most common user agents are graphical Web browsers.

Validator Software that tests code for syntactical conformity to a standard.

Value A specific setting chosen from a range of possibilities.

CROSS REFERENCES

See *Client/Server Computing; DHTML (Dynamic HyperText Markup Language); HTML/XHTML (HyperText Markup Language/Extensible HyperText Markup Language); XBRL (Extensible Business Reporting Language): Business Reporting with XML.*

REFERENCES

Alvestrand, H. (2001). *Tags for the identification of languages [RFC 3066]*. Retrieved June 22, 2002, from ftp://ftp.isi.edu/in-notes/rfc3066.txt

Bos, B. (1999). *W3C core styles*. Retrieved June 7, 2002, from http://www.w3.org/StyleSheets/Core/

Bos, B., Lie, H., Lilley, C., & Jacobs, I. (1998). *Cascading style sheets, level 2*. Retrieved May 20, 2002, from http://www.w3.org/TR/REC-CSS2

Clark, J. (n.d.). *SP*. Retrieved May 14, 2002, from http://www.jclark.com/sp/

Höhrmann, B. (n.d.). *W3C validator FAQ*. Retrieved May 13, 2002, from http://www.websitedev.de/css/validator-faq.html

HTML tidy (n.d.). Retrieved June 15, 2002, from http://tidy.sourceforge.net/

Lie, H. W., & Bos, B. (1999a). *Cascading style sheets: Designing for the web* (2nd ed.). Harlow, UK: Addison–Wesley.

Lie, H. W., & Bos, B. (1999b). *Cascading style sheets, level 1*. Retrieved May 20, 2002, from http://www.w3.org/TR/REC-CSS1

Ludwin, D. (2002). *A backward compatible style sheet switcher*. Retrieved February 10, 2002, from http://www.alistapart.com/issues/136/

Meyer, E. (2001a). *Complexspiral demo*. Retrieved May 14, 2002, from http://www.meyerweb.com/eric/css/edge/complexspiral/demo.html

Meyer, E. (2001b). *Liberty! Equality! Validity!* Retrieved February 4, 2003, from http://devedge.netscape.com/viewsource/2001/validate/

Meyer, E. (2002a). *Images, tables, and mysterious gaps*. Retrieved February 4, 2003, from http://devedge.netscape.com/viewsource/2002/img-table/

Meyer, E. (2002b). *CSS: Going to print*. Retrieved May 13, 2002, from http://www.alistapart.com/stories/goingtoprint/

Quinn, L. (2002). *WDG HTML validator*. Retrieved June 29, 2002, from http://www.htmlhelp.com/tools/validator/

Raggett, D., Le Hors, A., & Jacobs, I. (1999). *HTML 4.01 specification*. Retrieved May 17, 2002, from http://www.w3.org/TR/html401/

Sowden, P. (2001). *Alternative style: Working with alternate style sheets*. Retrieved November 2, 2001, from http://www.alistapart.com/issues/126/

Web Standards Project (n.d.a). *Browser upgrade campaign*. Retrieved June 20, 2002, from http://www.webstandards.org/act/campaign/buc/

Web Standards Project (n.d.b). *Browser upgrade campaign: Tips*. Retrieved June 29, 2002, from http://www.webstandards.org/act/campaign/buc/tips.html

World Wide Web Consortium (1999b). *Web content accessibility guidelines 1.0*. Retrieved June 29, 2002, from http://www.w3.org/TR/WCAG10/

World Wide Web Consortium (2001). *HTML validation service*. Retrieved June 29, 2002, from http://validator.w3.org/

World Wide Web Consortium (2002a). *W3C CSS validation service*. Retrieved June 29, 2002, from http://jigsaw.w3.org/css-validator/

World Wide Web Consortium (2002b). *Cascading style sheets, level 2, revision 1*. Retrieved September 12, 2002, from http://www.w3.org/TR/2002/WD-CSS21–20020802/

Zeldman, J. (2001). *A web designer's journey*. Retrieved March 14, 2001, from http://www.alistapart.com/stories/journey/5.html

C/C++

Mario Giannini, *Code Fighter, Inc., and Columbia University*

WHAT IS C/C++?

C and C++ are closely related programming languages used to create computer programs. Using these languages, programmers write documents called *source files* and then use a program called a *compiler* to convert the source files into program files that can be executed by a computer.

According to CareerInfo.net (http://www.acinet.org), computer software engineers are projected to be the fastest growing occupation, rising from 380,000 jobs in 2000 to an estimated 760,100 by 2010. The growing demand for technologically skilled software developers makes it a popular field.

The programs developed using C and C++ can cover a wide range types. Games, office products, utilities, server programs—just about anything can be developed using the C and C++ languages. C and C++ are not considered "high-level" programming languages, and unlike other languages such as COBOL or Visual Basic, they are designed to apply to a number of varied tasks, rather than specific ones.

A History of C and C++

The C language was originally designed by Dennis Ritchie for the DEC PDP-11 computer system and the UNIX operating system, between 1969 and 1973. It was directly influenced by the BCPL and B programming languages.

In 1969, Bell Telephone Laboratories was in the process of abandoning its Multics operating system, believing that its strategy would prove too expensive to completely implement. Ken Thompson led a group to investigate the creation of a new operating system, which would eventually evolve into the UNIX operating system.

Thompson wrote the original version of UNIX using assembly language, a complex and low-level language. He decided that UNIX would need a "system language," a programming language that would serve as the primary language for developing applications. Initially, he created his own language, named B. Thompson used BCPL (created in the mid 1960s by Martin Richards) as his inspiration for creating B.

B underwent several changes to improve several short comings and was renamed to the NB language (for, "New B"). As more changes were added, however, the language started to change considerably from its B predecessor. By 1973, the language was essentially completed and had been named C.

At the time, there was no "official" definition of C. The language had not undergone any standardization approval and was therefore open to interpretation by any company that wished to produce a C compiler. In 1978, Brian Kernighan and Dennis Ritchie wrote *The C Programming Language*. This was the most detailed description of the language and would soon be used to describe a version of the language named K&R C, for the author's names.

In 1983, the documentation from Kernighan and Ritchie's book was submitted to the American National Standards Institute (ANSI), to get their approval for an exact and standard definition of exactly what C was. Once approved, then the C language would carry with it the extra clout of having an official standard, to which compiler producers would need to meet in order to sell their products as an official C compiler. By 1989, after several changes, the ANSI XJ311 committee approved what is now called ANSI C. This is the language still in use today.

While C was undergoing its many changes for standardization, C++ was appeared around 1980 and continued to grow in popularity. C++ was originally termed "C With Classes" and was created by Bjarne Stroustrup at AT&T Bell Labs. Stroustrup was using Simula67, which followed an object-oriented approach to software development. Unpleased with Simula67's performance, Stroustrup decided to enhance the C language, to implement classes and objects.

Since its introduction, C++ has been considered a superset of C; everything in C is pretty much still in C++. C++ adds a number of features, primarily in the area of object-oriented programming. It is still possible to compile a C program using a C++ compiler, but not vice versa. The name C++ is a take on the ++ operator of C, which means "add one." So, C++ can be considered as "C + 1," or the next evolution of C.

Stroustrup created C++ in an informal fashion. There was no formal committee or project dedicated to its creation. It more or less promoted itself into development, offering the performance of C++, but with object-oriented enhancements and patterns to improve development. Because of this informal process, there is not a great deal of information about the evolution of its specification. Despite this fact, and as with the early years of C with no standard definition, C++ grew quickly in popularity. By 1987, it was decided that C++ should be submitted to the ANSI body for official standardization. By 1989, Stroustrup had completed the documentation required for standardization, and Hewlett Packard had initiated the process for ANSI submission.

In 1989, ANSI created the X3J16 committee, which was to serve as the review body for standardization approval. During the review process, C++ underwent a number of changes. An official standard stream (or Input–Output) class hierarchy was defined, as was a set of standard "template" classes. By 1995, a proposal was made that would eventually be approved. By 1998, ANSI had approved and published a C++ standardization, which is the language still in use today.

Programming Languages

Everything a computer does is the result of a program that runs on it. Even when a computer seems to be idle, waiting or a user to press a key, it is in fact doing many things. A program is defined as a set of instructions that the computer executes. A computer follows an ordered pattern of instructions, just as a person reads a recipe to bake a cake and executes the recipe's instructions in order.

Each computer (or central processing unit [CPU] inside the computer) understands a specific set of instructions. The Intel CPUs found inside Windows computers executes a different instruction set then Motorola CPUs found inside an Apple Macintosh computer. The instructions the CPU knows how to handle are called its *instruction set*. Even thought the Intel and Motorola CPUs have a different instruction set, many of the instructions are similar. For example, both provide instructions to do simple math such as addition and subtraction, as well as to move data from one memory location to another or compare the results of some mathematical operation. The instructions executed by the CPU are also called *Machine Code,* because the machine can execute it directly.

A CPU's instruction set represents everything it understands and can execute. Unfortunately, the instruction set is defined as a set of numbers, and each number tells it to perform a certain operation. This makes creating a program using the exact instructions quite complex. For example, the following numbers (in hexadecimal) on an Intel CPU would tell it to add "1" to memory address "100":

```
C6 06 64 00 01
```

Using this format can become complicated, given that each instruction performs only a small part of a program's function (even a very small program, such as Windows' Notepad program, contains about 50,000 bytes of instructions). To simplify the creation of a program, machine code instructions need to be made less complex. The first step in achieving this is to introduce a set of mnemonics for the instructions. These mnemonics are a representation of the machine code that is more readily understood by humans, called *Assembly* or *Assembly Language*. An example of the same set of instructions, written in assembly, would look like this:

```
ADD BYTEPTR[100],1
```

Now, by just looking at the mnemonics, one can begin to understand that the program is attempting to add 1 to a memory address (memory address 100). The computer doesn't understand mnemonics, however; it understands the numbers in its instruction set. Therefore, an *Assembler* program is used to convert Assembly Language into actual CPU instructions.

Assembly is defined as a low-level language. This means that it offers programmers complete control over the programming process, but they must take responsibility for every operation, no matter how minor. Higher level languages were developed, like C and C++, to simplify the task of creating programs.

A single statement or instruction in C or C++ may translate into several machine code instructions. For example, the following code adds "1" to a variable in C or C++:

```
Count = Count + 1;
```

This is much easier to read and understand than the assembler mnemonics.

Compilers and Linkers

A compiler is a program that translates instructions and statements from one language, such as Fortran or C, into machine code (code that the computer can directly execute). The output of the compilation process is called an object module and is made up of the translated machine code. A linker takes one or more of these object modules and links them together into a single file called an executable. The idea behind linking is that if a program has a set of instructions to do something useful, such as printing a string, the programmer will want to be able to reuse that set of instructions in this and other programs. If the programmer compiled these instructions into an object module, then he or she simply needs to link it into a new program to reuse it.

All programming languages are defined as either *interpreted* or *compiled.* Interpreted means that the program is not directly executed by the computer, but by a program on the computer. Compiled languages are ones that are compiled (or translated) directly into machine code, which can run directly on the computer. They are typically faster than interpreted programs, but interpreted programs are generally easier to create and debug. PERL, Java (with the JVM), and PHP are all examples of interpreted languages. C and C++ are examples of compiled languages.

To simplify programming, developers can write their programs using C language, following its rules and syntax. They then compile the program and link it to whichever object modules they need to create an executable program file. The computer is then able to execute that file directly.

To create an executable program using the C and C++ languages, programmers need the following items:

- A text editor, to type in the C or C++ statements
- A compiler, to convert the C or C++ statements into machine code (object modules)
- A linker, to join one or more object modules into an executable file

Most operating systems come with some form of text editor, such as *vi* or *Emacs* in Unix or *Notepad* in Windows. Unix and Linux operating systems also come with a built-in C and C++ compiler and linker, but Windows and MacOS do not. When developing programs, however, most programmers will use an integrated development environment (IDE). These are programs are actually a suite of programs that combine text editor, compiler, linker, and other helpful tools such as a debugger into a single application. For Windows, common IDEs include Visual C++ from Microsoft and C++ Builder from Borland. For the Macintosh, CodeWarrior from Metrowerks sells a C an C++ IDE.

C Versus C++

As noted earlier, C++ is often considered a superset of C; everything C can do, C++ can do as well, but not vice versa. With a C++ IDE, one can still create "pure C" programs (programs that do not take advantage of the benefits of C++), but C++ adds several useful attributes to C, including the following:

- C++ introduces classes and a syntax to create and use them. Classes are a syntactical method by which the programmer can group data and functions into a simple package or container.
- C++ introduces a set of predefined classes and functions called the standard template library (STL) to reduce development time. STL increases functionality, for example, by offering a means to manage data collections and sort or search those collections.
- C++ introduces function overloading (the ability to write several functions with the same name). This permits a more intuitive means for naming functions to manipulate different types of data for similar tasks (e.g., three "clear" functions that clear three different entities—a string, a file, or a list of data).

C++ provides developers a way to make smaller, more manageable "objects" to work with data in their programs, rather that trying to string together a set of similar but unrelated functions. This, combined with a string library of classes such as STL, permits faster program development, a key benefit of C++.

The remainder of this chapter provides examples in C++. Some of these will work in a "Pure C" program, but others will not. In certain cases, such as printing a string to a console, each language has its own way of performing a task, but only the C++ version is given here (except where noted).

GETTING STARTED WITH C/C++

I begin by demonstrating a simple C and C++ program that will display the text "Hello World" on the console (screen) and then terminate. Although relatively simple, upon completion you will have written a small program and compiled, linked, and executed it.

First, create a *source file* to contain the program and its C++ statements. Depending on the compiler, a programmer would normally add a .cpp extension to the file name, although some compilers prefer .cxx or some other variation. Here, assume that the source file can be called helloworld.cpp. Place following code in the source file:

```
//Include header file so we can use the
  cout object
#include <iostream>

//This next line simplifies using the std
  STL library classes and
//objects, like cout:
using namespace std;

/* main is where the program starts */
int main()
{
        //Use the cout object to display a
          string to the console.
        cout << "Hello World" << endl;
        return 0;
}
```

Now that this text has been added to your helloworld.cpp file, you will compile, link, and execute it (the exact steps depend on the compiler and linker you are using, as well as your operating system). The program demonstrates several issues, discussed below.

Comments

Comments in C/C++ are pieces of text that the compiler ignores. They are used to permit the programmer to make notes and comments about the source code. C permits developers to put comments (as many lines as he or she desires) between the /* and */ characters. C++ adds a // comment sequence, where anything following these two characters, to the end of the line, is treated as a comment.

Functions

A function is a group of C/C++ statements that can be called on from various places within a program. A function is defined as a return type, then the function name, and a list of parameters within parenthesis, and finally the actual set of C or C++ statements enclosed in curly braces {and}. In the previous example, *main* is a function that returns an int data type and accepts no parameters. It also has special significance, because this is where a C or C++ program begins execution. Sometimes main is written to take two parameters, commonly named argc and argv. In

Table 1 Primitive Data Types for C and C++

TYPE	DESCRIPTION	EXAMPLE
char	Represents a single character	char Ch = "A"
short	Represents a small integer, usually 16 bits in size, that can contain a value from –32768 to 32767 (or 0 to 65536)	short S = 0;
long	Represents a large integer, usually 32 bits in size, that can contain a value between –2 billion and 2 billion (or 0 and 4 billion)	long l = 0
int	Represents an integer that typically defaults to a *long,* but can also be a short	int i = 0
float	Represents a floating point (non–whole number) value, typically with seven digits of precision	float f = 3.142857
double	A floating point (non–whole number) value, typically with 15 digits of precision	double d = 3.14285714285714

this version, argc will contain the count of command-line arguments passed to the program, and argv will be a string array of those arguments.

Let's say I decide a useful function would be to ask the user to input a number. Rather than repeating code to perform this task, I place it inside a function and simply call or invoke the function as needed. For example:

```
int PromptInt( const char* Prompt )
{
      /* This function prompts the user for
         input, and returns the input */
      int Ret;
      cout << Prompt << endl;
      cin >> Ret;
      return Ret;
      /* The return statement permits a
         function to return a value to its
         caller (main, in this example) */
}
int main()
{
      //Use the cout object to display a
         string to the console.
      cout << "Hello World" << endl;

      int Age = PromptInt
         ( "Enter your Age: " );
      int Weight = PromptInt
         ( "Enter your weight: " );
      return 0;
}
```

Note how the actual number of steps to output a prompt (e.g., "Enter your Age:") and accept an input from the user takes four lines of code. I have placed those 4 lines of code into the *PromptInt* function, however, so that when I need to do this task I only have to write one line of code to call the *promptInt* function, such as in the line:

```
      int Age = PromptInt
         ( "Enter your Age: " );
```

Creating functions allows programmers to reuse code easily and therefore simplifies the task of programming. Some other languages refer to a function with a return value as a "function" and one without a return value as a "subroutine"; C and C++ do not. The term "method" is often used to describe a function that exists within a class.

Simple Data Types

A program is made up not only of coded instructions but of data as well. For example, when people use a word-processing program, the text they type is stored as data in memory while the program is running. This data is typically stored to a disk, so that when the user shuts down the program, the data (i.e., the document) is retained for future use. When the word processor is run again, it simply reloads the file into memory.

C and C++ provide several basic data types to do various things. These are all called fundamental or primitive data types because they are relatively simple. They can be used to construct more complex data types, however, such as structures or classes. The core list of primitive data types for C/C++ is shown in Table 1. These data types can be used in a variety of ways. For example, a short might be used to count from 1 to 100, to execute something 100 times, or a float might be used to hold a person's hourly pay rate or a foreign currency exchange rate.

Pointers

Pointers are often viewed as the most complex part of learning C and C++, but they are actually simple. A pointer is a variable that contains the memory address of some other variable or function. Pointers, like all variable types, require proper initialization. C and C++ are sufficiently high level that developers need not worry about the exact address.

Pointer variables are declared similarly to other variables, bt they use the * operator in their declaration. For example:

```
int i; /* i is an integer variable */
int * ip; /* ip is a pointer to an
  integer */
```

Initializing a pointer can be done a number of ways, but a fairly good example is to use the "address of" operator &, with an existing variable:

```
ip = &i; /* Initialize ip with the
  'address of' i */
```

Now, I again use the * operator to dereference the data to which ip points (i.e., of which ip contains the memory address). For example:

```
*ip = 3; /* Puts 3 into the i variable */
```

Note the difference between `ip = &i;`, which initializes ip itself, and `*ip = 3;`, which initializes the integer to which ip points.

Arrays

An array is a set of variables in contiguous or continuous order. Although declaring a single integer data type to hold a value like a counter is important, practical programming often requires an array of integers to hold several things, such as students' IDs. An array is declared by placing brackets ([]) and the number of desired elements after the variable name. For example:

```
int StudentIDs[30];
```

This example creates an array named StudentIDs that contains 30 integers. In C and C++, access to an individual element of an array starts at zero, so the first integer in the array is named StudentIDs[0] and the last is StudentIDs[29]. Once an array has been declared, an individual element can be treated like any other variable of that same type (integers, in our example). For example:

```
int Index = 0;
cout << "Enter student ID: " << endl;
cin >> StudentIDs[Index];
```

Arrays can be multidimensional. For example, consider the following:

```
int Month[5][7];
```

Here, month is a two-dimensional array, for which one can imagine 5 rows of 7 days (one wouldn't use 4 rows, because that would only be 28 days).

Simple Input and Output

C++ has two objects to permit programmers to input (get input from the keyboard) and output (display something on the screen), named *cin* and *cout* respectively. These objects are actually global variables that are predefined. Both cin and cout can work with the various fundamental data types, meaning that cin knows how to get integers ('int') and strings (char arrays or *string* objects) from the keyboard, and cout knows how to format the same types for display.

To display or output a value, programmers use the cout class, and the << (insertion) operator. cout can work with variables or constants (as in the constant string example). cin can only work with variables, because it stores data into them. Using cin requires use of the >>>> (extraction) operator. For example:

```
#include <iostream>
#include <string>
using namespace std;//Simplifies usage of
  cin, cout, and string

int main()
{
       string Name;
       int Age;

       cout << "Enter your name:" << endl;
       /* endl is an End of Line character,
         and a buffer 'flush' */
       cin >> Name;
       cout << "Enter your age:" << endl;
       cin >> Age;
       cout << "Hello " << Name << ", you
         are " << Age << " years old."
         << endl;

       return 0;
}
```

In this code, the string *Enter your name:* is output by cout, and then the user types his or her name and presses ENTER, which cin takes and places into the Name variable. A similar process is performed for the *Age* variable. An example program run would look like this (users input are in **bold**):

```
Enter your name:
Mario
Enter your age:
38
Hello Mario, you are 38 years old.
```

C provides printf and scanf functions (among several others) for performing output and input, respectively. Additional details are beyond the scope of this chapter, but it is worth noting that the same routines will work in C++ programs as well.

Classes: An Introduction

As described in the previous section, C and C++ programs comprise functions or subroutines that operate on data or variables. It would be beneficial to take functions that operate on a certain type of variable and combine them with the variables on which they work. This is just one of the things that a *class* can do.

The definition of a class is covered in more detail later in the chapter and is closely related to object-oriented programming. For the purposes of this section, however, a class is a simple means of grouping functions and variables together. In other words, a class is used as a container that holds both data and functions that work on that data.

Imagine that I have a person's name and age in variables, and I have functions to input and output the person's data. By using a class, I can combine them into one object. Creating a class in C++ represents declaring a new data type using the *class* keyword. The following is an example of a class definition:

```
#include <iostream>
#include <string>
using namespace std; //Simplifies usage of
  cin, cout, and string

class Person {
public:
       string Name;
       int Age;
       void Input() {
              cout << "Enter person name:"
                << endl;
              cin >> Name;
              cout << "Enter person age:"
                << endl;
              cin >> Age;
       }

       void Output() {
              cout << "Name: " << Name
                << " Age: " << Age << endl;
       }
};

int main(int argc, char* argv[])
{
       Person p; //p is now a Person object,
         with a Name and Age inside it.

       p.Input();
       p.Output();

       return 0;
}
```

This program's output would look like this (user input in bold):

```
Enter person name:
Mario
Enter person age:
38
Name: Mario Age: 38
```

The *Person* class I created contains two variables: Name and Age. Each instance of a Person object I create, such as *p*, will contain its own Name and Age variable. Because functions never change, there is only one copy of each function in the class, no matter how many instances of a Person object we create.

FLOW CONTROL

Flow control refers to how a program executes or what path it takes through its lines of code as it executes. Flow normally goes sequentially from line to line, executing each line of a function in top-down order. There are times when programmers may want to repeat certain sections of code or to skip other sections altogether. To accommodate this, C and C++, like all other languages, use flow control.

Expressions

As mentioned earlier, programmers may want to execute a certain piece of code multiple times or bypass it all together. To determine if they should repeat the code of skip it, programmers rely on an *expression*. Expressions can be logical, which are either true or false or mathematical, which result in a number (zero is considered as false, and anything else is considered true). In C and C++, programmers can use any type of expression to determine flow control, although logical expressions are the most common. For example, when inputting a person's age, I may decide that if the user enters a bad Age (such as –6 or 1,000), that the program will prompt for the age to be reentered. My expression would be *Age < 1 || Age > 120*. This expression says "Age is less than 1 OR Age > 120" (The || means 'or' in C and C++). To apply this in an actual example, we might see something like the following:

```
     do {
            cout << "Enter person age:"
              << endl;
            cin >> Age;
     } while( Age < 1 || Age > 130 );
```

This code uses a looping flow control statement called a *do while*. It will "do" the code in between the {and} braces, while Age < 1 or Age > 130 (under the assumption that people don't live past 130 years).

For an OR expression to be true, at least one of its sides must evaluate to true on its own. There is also an "and" operation, which is written with the && symbols. For an AND expression to be true, both of its sides must evaluate to true. I can mix and match the || and && operators as needed. For example, the following code will tell me if the variable Ch contains a digit from 0 to 9 or a + or – character:

```
if( (Ch >='0' && Ch <='9') || Ch=='-' ||
  Ch == '+')
```

Note that in this example, > = means greater than or equal to, < = means less than or equal to, and = = means equal to. You can also use ! =, which means not equal to.

Looping and Nonlooping

Flow control statements are divided into two types: looping and nonlooping. As the names imply, looping statements can cause a block of code to be executed multiple times, and nonlooping statements will permit the programmer to selective execute a piece of code only once.

The nonlooping flow control statements are *if*, *switch*, and *goto*. Of these, the *goto* does not use an expression test to determine if it should go the specified point. *goto* statements are usually overused when a programmer doesn't comprehend the other flow control statements, so many new programmers are warned not to use them.

If Statement

The *if* statement has an optional *else* that permits programmers to execute alternate pieces of code. The following are three variations of using the if and else statements:

Version 1

```
if(Age >= 65)
      cout << "Eligible for Social
        Security" << endl;
```

Version 2

```
if(Age >= 65)
      cout << "Eligible for Social
        Security" << endl;
else
      cout << "Not eligible for Social
        Security" << endl;
```

Version 3

```
if(Age >= 65)
      cout << "Eligible for Social
        Security" << endl;
else
{
      if(Age < 18)
            cout << "Not eligible to work"
              << endl;
      else
            cout << "Eligible to work"
              << endl;
}
```

Switch Statement

The switch statement performs similarly to an if–else combination, although it has some differences. The switch optionally permits multiple code sections to execute, depending on the order and structure of the code. For example:

```
int Vowels=0, Letters=0, Spaces=0;
char Ch='a';

switch(Ch)
{
       case ' ':
              Spaces = Spaces + 1;
              break; //Take us out of this
                switch
       case 'a':
       case 'e':
       case 'i':
       case 'o':
       case 'u':
              Vowels = Vowels + 1;
              //No break, we fall through to
                the next line of code
       default:
              Letters = Letters + 1;
}
```

This code will add 1 to the Spaces count if Ch is a ‘ ‘ character, 1 to both the Vowels and Letters count (vowels are letters) if Ch is a vowel, or 1 to the Letters count if Ch is not a space or vowel.

For Statement

The for statement combines an initialization section and an operation section, along with the expression test to determine if it should repeat. The basic format is:
for(initialization; Expression Test; Operation)

A practical example of this is the following:

```
void PrintStars( int Count )
{
       int i;
       for( i=0; i < Count; i = i + 1 )
              cout << "*";
       cout << endl;
}
```

In this example, the *initialization* sets the variable *i* to zero. The *Expression Test* would be to see if *i* were less than Count (if it is, the body executes; if it isn’t, the body is skipped and the program continues). If the body was executed, then the *Operation* is performed, which in this case adds 1 to *i*. The *Expression Test* is then reevaluated, and the process repeats.

While and Do While Statements

The *while* statement works basically like the *for* statement, only it has no specific initializer or operation steps. More often than not however, they are found in or near the *while* statement. The format is as follows:
while(Expression Test)

An example would be:

```
void PrintStars( int Count )
{
       int i;
       i=0; //Similar to initializer in
         for loop
       while( i < 10 )
       {
             cout << "*";
             i = i + 1; //similar to
               operation in for loop
       }
       cout << endl;
}
```

The difference between the *while* and *do while* is that the expression test occurs at the end of the body, and the body is guaranteed to execute at least once. An example of a *do while* loop can be found at the start of this section.

Exceptions

C++ introduces the concept of an exception, which C does not support (there is a roughly similar technique in C++ using the *setjmp* and *longjmp* functions). An exception is a way to jump from the middle of a function back to the middle of another function that may have called it. Many programmers think of exceptions as useful only for undoing errors, but although this is their most common use, they can be used at any time to leave one section of code and return through several function calls. A lengthy description of exceptions is beyond the scope of this chapter, but the *try, catch,* and *throw* key words are used to implement exceptions.

ADVANCED DATA TYPES

C and C++ provide several more advanced and complex data types, as well as the ability for programmers to create their own. As noted earlier, the *class* key word can be used to create a simple container; following is a more detailed description of that process, as well as other advanced data types.

Structures and Unions

In C++, a structure works just like a class does, except that members are public by default, whereas in a class they are private by default. Like classes, a structure in C++ can also contain member functions. In C, structures can only contain data members, and no functions.

A typical structure definition uses the *struct* keyword, and can be demonstrated with the following:

```
struct MyDate {
       int Month, Day, Year;
}
```

Here, MyDate is a new type of data that contains three integers: Month, Day, and Year. In memory, each data member of the struct is laid out sequentially. This means that the size of the struct in memory would be the size of three integers.

A union has a similar syntax as a struct but has a significant difference: Each member of a union occupies the same memory address. As a result, the size of a union in memory is the size of its largest member. The following is an example:

```
union DateOrTime {
       struct { int Month, Day, Year;} Date;
       struct { int Hour, Minute, Second;}
         Time;
};
```

The DateOrTime union has a Date and a Time data member, but these members overlap each other. This means it can only contain a valid Date or a valid Time, but not both. The size of the union would be the size of three integers.

To access a data member of a struct or union, the programmer must first have a variable of that type and then use the operator (or the -> operator if you have a pointer to a struct or union). For example:

```
struct MyDate BirthDay;
cout << "Enter Month: " << endl;
cin >> MyDate.Month; /* Get integer from
  user, and store it in a data member */
```

Classes

Classes can be used for more than a simple container. There are a wide variety of topics to address in classes, but I cover only two issues: constructors and destructors and inheritance. Accessing the member variables or methods of a class, is similar to that of the struct or union, and uses the. operator (or the -> operator if you have a pointer to a struct, union, or class).

Constructors and Destructors

A class can provide two special functions, called a constructor and a destructor. These functions are automatically called when an object is created or destroyed, respectively (objects are typically created when they are declared as a variable and destroyed when the function that declared them is done running).

Both the constructor and destructor functions have no return value and must have the same name as the class they are in, but a destructor must have a tilde (~) prefix in its name as well. Constructors are used to initialize the data members of a class, whereas a destructor is used to clean up any resources (such as memory or an open file) that the object may hold. An example of a class with a constructor would be as follows:

```
class Person {
public:
       string Name;
       int Age;

       Person () // Constructor, to
         initialize data members
       {
              Age = 0;
              Name = " ";
       }

       void Input()
       {
              cout << "Enter person name:"
                << endl;
              cin >> Name;
              do {
                     cout << "Enter person
                       age:" << endl;
                     cin >> Age;
              } while( Age < 1 || Age > 130 );
       }

       void Output()
       {
              cout << "Name: " << Name <<
                " Age: " << Age << endl;
       }
};
```

Note that this class has no destructor. If the class had been designed so that it opened a file or perhaps allocated some other type of resource, then a destructor would have been written to ensure that the resource was closed when the object was being destroyed.

Inheritance

A class (and structure and unions as well) can be used as the base class, or starting point, for another class. When programmers do this, they create a new class that inherits all the data members and functions of the base class. This permits them to reuse classes and build on them without having to modify the original version.

The following example demonstrates a simple class derived from the Person class:

```
class Worker : public Person
{
public:
             int Salary;
             void Input()
             {
                  Person::Input();
                    //Get person Input
                  cout << "Enter Worker
                    Salary: " << endl;
                  cin >> Salary;
             }
             void Output()
             {
                  cout << "Name: " << Name
                    << " Age: " << Age;
                  cout << " Salary: "
                    << Salary << endl;
             }
};
```

If I were to declare a Worker object, such as `Worker w;`, the *w* variable would contain a Salary integer (from the Worker class) and an Age and Name variable that it inherited from the Person class. By simply adding "public Person" to the declaration of the Worker class, I am saying that it is derived from the Person class.

The Standard Template Library

The ANSI standard version of C++ provides an entire collection of classes and functions, referred to collectively as the standard template library. Inside the library are classes and methods required to do common programming tasks. They implement collection classes for collecting and sorting data in a variety of fashions such as arrays, lists, and maps. The STL also provides classes for string management, file and stream input–output, as well as support to legacy C functions.

C AND C++ AND THE INTERNET

As mentioned at the beginning of this chapter, C and C++ are very generic languages, suited to serve a wide range of purposes. They are used to create everything from games, to device controllers, to desktop and server applications. I also noted that C and C++ can be used to create operating systems such as Unix and Windows, on which all other programs are run. This puts C and C++ in a unique position. Many other languages such as Practical Extraction and Report Language (PERL), Java, PHP: Hypertext Preprocessor (PHP), JavaServer Pages (JSP), or Active Server Pages (ASP) share a well deserved popularity, the truth is that most, if not all of these languages were actually created with C and/or C++ to begin with.

As a typical example, the Unix environment and the Apache Web server application are used to run many of the Internet's Web sites. At the time of this writing, the Apache Web server accounted for 59% of all sites surveyed (http://www.netcraft.com), and it was written in C. In addition, the scripting language PERL is commonly used to create CGI programs; PERL (or, interpreter) also was written in C. Put simply, C truly is the building block of the Internet. Although other languages may be more popular for Web site scripting, they wouldn't exist without C.

CGI Programming

A common gateway interface (CGI) program is one that is run on a Web Server, the output of which usually goes back to a clients' browser programs. Just about any time users fill in a form on a Web page that has a "Submit" button, it is gathering data from the form and sending it to a CGI program on the remote Web site.

Any programming language that can display output to the console can be used to create a CGI program, and this includes C and C++. If the program can read data from the keyboard or from an environment variable, it can also support form-based input from a clients browser. Again, C and C++ fit this bill. Because C is a compiled language and not an interpreted language like Java, Perl, or ASP, CGI programs created in C or C++ typically perform tasks faster than similar programs written in other languages.

How a CGI Program Is Run

A Web page typically uses a Hypertext Markup Language (HTML) FORM tag (which defines a set of controls such as edit boxes and drop-down lists for user input) to define a CGI program on the Web Server to be called. An example of such an HTML tag is the following:

```
<FORM ACTION= ”CGIFavColor.exe”
  METHOD=POST>
```

Here, this form should run the CGIFavColor.exe CGI program, with a POST method. CGIFavColor.exe would be a compiled C program on the Web server. Note that running the CGI program successfully depends on specific needs of different servers; for example, a programmer might need to set certain permissions for the CGI file and require a certain extension. In the example, using the .exe file extension assumes that the CGI program resides on a Windows-based Web server, and the same extension would not be used for a Unix-based CGI program.

Most commonly, a "Submit" button is used to invoke the CGI. A submit button is a special HTML tag intended for just this purpose. The HTML code for this submit button would look like this:

```
<INPUT NAME="RunTheCGI" TYPE="submit"
  VALUE="Submit Color">
```

The submit button will ask the Web Server to invoke the CGI program specified by the FORM tags *action* property. It sends all the data from the form (entered via the client's browser) to the server.

There can also be a link on a page run by a CGI program. To do this, the programmer specifies the name of the CGI program as the link, for example (this example also shows how to pass a parameter to the CGI):

```
<A HREF="../cgi-bin/CGIFavColor.
  exe?FavColor=Black">Favorite Color is
  black</A>
```

The link method does not use (or need) the FORM tag to specify a CGI program to run.

A CGI program can also be run when a page is loaded (often used for Web counters) by specifying the name of the CGI as an element, like an image. For example:

```
<IMG SRC="../cgi-bin/CGIFavColor.exe">
```

This method would assume that the CGI program would output an image, like a GIF file. In this case, when the tag is encountered by the browser, the CGI program is run, and it can do whatever it wants (e.g., increment a counter in a file), and then it returns (outputs) a GIF image.

CGI Output

A CGI program outputs its result by sending the data to the standard output device (using stdout, or cout in C++). The programmer must format all output in a specific format, however. The most common format is HTML text, but a program can also output an image, sound, or any other file type that a browser will recognize. For the remainder of this section, I describe HTML format, because it is by far the most common (an exception being access counters, which often output an image).

A CGI program must begin its output (again, for HTML) with the string `Content-type: text/html`, followed by two carriage returns. In C, this would be accomplished with the following:

```
printf( "Content-type: text/html\n\n");
```

In C++, we could do the following:

```
cout << "Content-type: text/html\n\n";
```

Next, the CGI must output the entire Web page using HTML tags. The minimum required would be the start and end BODY and HTML codes. The programmer can insert anything he or she wants between the start and end codes. For example:

```
printf("Content-type: text/html\n\n");
printf("<HTML><BODY>\n"); /* \n isn't
  needed, but makes source more readable */
printf("Hey! This is my first CGI
  response!<P>\n");
/* <P> is an HTML paragraph tag */
printf("</HTML></BODY>\n");
```

The Web server that ran the CGI program gathers all the output from the *printf* function call, and then redirects it back the client's browser for display.

Note that even error messages from the program must appear in the format.

CGI Input

CGI input, typically the user's data entered in HTML, is presented to the CGI program in one of two methods, GET or POST. The HTML FORM tag defines the method to use. The difference between the GET and POST is how the information from the form is sent to the CGI program from the server, and how it must retrieve the data.

A GET method will provide the user's input to the CGI program as an environment variable called QUERY_STRING. The CGI program would read this environment variable (using the C getenv() function) and parse it to get the user's input. A GET method will also show the input data to the user in the URL area of the browser, showing a string such as http://www.somewhere.com/CGIFavColor.exe?FavColor = Black. The GET method is acceptable for small amounts of data. Also, GET is the default method when a CGI program is run via a link.

A POST will provide the user's input to the CGI program as if it were typed at the keyboard using a standard input device, or stdin. If POST is used, then an environment variable called CONTENT_LENGTH indicates how much data is being sent. The programmer can read this data into a buffer by doing something such as the following:

```
char Buffer[512];
int InputLength = atoi(getenv
  ("INPUT_LENGTH"));
fread(Buffer, InputLength, 1, stdin);
```

The CGI program should inspect the REQUEST_METHOD environment variable to determine if the form

Table 2 C and C++ Open Source Programs

PRODUCT	DESCRIPTION
Linux	Unix-style operating system
MySQL	Structured query language (SQL) database server
Apache	Web server
Mozilla	Web Browser
Doom Legacy	Original Doom arcade game
Tcl	An interpreted Tool Control Language
Firewall Builder	Graphical tool and compilers for various firewall platforms
Enlightenment	Window Manager for X11 windowing system
CDex	CD Burning software

was a GET or POST method and take the appropriate action to retrieve the data from the user's form.

Parsing CGI Input

Once the input is received, it must be parsed. All field values from the form will appear as one long string. Each value is separated by an & character. For example, a two-value string might look like this:

```
FavColor=Black&FavFood=Twinkies
```

Here, FavColor and FavFood would be the names of two input fields from the form, and Black and Twinkies would be what the user entered in those fields.

The server will also perform two conversions on the data the CGI program receives:

- All spaces are converted to a + character
- Special characters (such as \n) are converted to a %, followed by the ASCII code for the character

The programmer needs to undo these conversions in the CGI program. Once the FORM input data has been obtained, the programmer can then do whatever he or she wants with the data. Because C++ has the ability (with the proper code and functions) to send e-mails, manage databases, create files, and so on, programmers can combine any of these tasks in their CGI programs. For example, one could get the users e-mail address (if it was entered in an HTML form), add it to a database, and then send an e-mail confirming they were added.

CLIENT–SERVER PROGRAMMING

C and C++ are used more in client–server programming than as a CGI program. A typical example of a client server program would be Apache or Internet Information Server as a Web server and Netscape Navigator or Internet Explorer as a Web browser client.

C and C++ are often used to write programs for both the client and server. Programs that communicate with each other over the Internet do so using something called a *socket*. A socket isn't a physical thing but a description of how to exchange data with other computers and programs. Although C and C++ do not have built-in support for sockets, the operating system under which the programs run usually provide a set of functions to work with sockets, which can in turn be called by C and C++ programs.

Typically, a program that acts as a server is running constantly, waiting for a client program to connect via a socket. The client program establishes the connection to the server by calling the appropriate operating system function. Once connected, the programs can freely transfer data back and forth, such as HTML files, email messages, and files.

Open Source

The Open Source movement is the practice of developing software applications and then sharing the source code for those programs with others, free of charge. The basic license, called the GNU license (GNU is a recursive acronym for GNUs Not Unix), basically states that programmers are free to use the source code as long as they also distribute the source code of their programs under the GNU license.

One of the largest Web sites that functions as a server for Open Source codes is http://www.sourceforge.net. A quick glimpse at its top-10 downloads of all time (http://sourceforge.net/top/toplist.php?type = downloads) shows that 8 of the top-10 items were mostly written with C and C++ (Delphi and PHP are the remaining two).

Table 2 is a brief list of some of the more popular C and C++ Open Source programs. The list demonstrates that C and C++ are used for a wide range of applications.

GLOSSARY

American National Standards Institute (ANSI) An organization that defines recognized standards.

Common gateway interface (CGI) A program run on a Web Server, usually at the request of a client (browser).

Class A template that defines code and data to perform a particular task; a data type.

Client An application that connects to a server to use a service, such as sending or receiving e-mails, or displaying Web pages.

Compiler A program that converts a source file into computer instructions.

Executable program A program that may be executed directly by a computer.

Inheritance The ability to create a new class from a preexisting class and inherit its functionality.

Instruction set The instructions or directives that a particular computer can execute.

Interpreted language A programming language whose programs rely on a running application to be parsed and processed to execute; PERL and PHP are examples of interpreted languages.

Linker A program that joins compiled program modules into an executable file.

Server An application that waits for client programs to connect and provides them with a service, such as sending or receiving e-mails or sending Web pages to a browser client.

Socket A method by which two programs can exchange data on either the same or different computers; Client and server programs typically communicate with each other using a socket.

Source file A text file that contains program execution statements, as keyboarded by a programmer.

Standard template library (STL) A set of functions and classes in ANSI Standard C++ that provides various functionality.

UNIX An operating system developed by Bell Labs.

CROSS REFERENCES

See *Client/Server Computing; Common Gateway Interface (CGI) Scripts; Linux Operating System; Open Source Development and Licensing; Unix Operating System.*

FURTHER READING

This chapter is intended to provide only a brief summary of the C and C++ languages. For further information, reader may refer to the following books and Web pages.

Eckel, B. *Thinking in C++*. Retrieved from http://www.mindview.net/Books/TICPP/ThinkingInCPP2e.html

Josuttis, N. M. (1999). *The C++ Standard library: A tutorial and reference*. Addison-Wesley.

Kernighan, B., & Ritchie, D. (1988). *The C programming language*. Prentice Hall.

Koenig, A., & Moo, B. E. (2000). *Accelerated C++: Practical programming by example*. Addison-Wesley.

Lippman, S. B. (1999). *Essential C++*. Addison-Wesley.

Online resource for C and C++ Programming. Retrieved from http://www.cprogramming.com

Stroustrup, B. (1995). *The C++ programming language.* Addison-Wesley.

The Development of the C Language. Retrieved from http://cm.bell-labs.com/cm/cs/who/dmr/chist.html

Free C++ Compilers can be obtained from the following Web sites:

http://www.bloodshed.net/
http://www.borland.com
http://www.openwatcom.org/
http://developer.apple.com/tools/mpw-tools/
http://www.delorie.com/djgpp/

Additional online resources in the form of compilers, libraries, and documentation can also be found at the following sites:

http://www.thefreecountry.com/developercity/index.shtml
http://www.openroad.org (Academic C and C++ programming)

Circuit, Message, and Packet Switching

Robert H. Greenfield, *Computer Consulting*

INTRODUCTION

Circuit, message, and packet switching are techniques for transferring information. These concepts are not unique to electronic networking. They have common, everyday models and historical prototypes. A common example of circuit switching is a voice telephone conversation between two people. Message switching is seen every day in the postal, paper-based mail system. Visualizing packet switching takes more imagination. Let's move a complete business—staff, furniture, files, business machines, and whatever—from its old location to a new building. We assume that the business is sufficiently large that it needs several autos and buses to move the people and a number of trucks to move the nonhuman assets. All these vehicles, each containing a portion of the company, move over a system of roads. Together, as an aggregate, the company is the sum of the payloads of all the vehicles.

We can take this "company-moving" example a little further to illustrate datagrams and virtual circuits. If we allow all the cars, trucks, and buses to select their own routes from the source to the destination, we have a datagram example. Each vehicle arrives on its own schedule. However, if we dictate a specific route, mandate a convoy, or a parade, by assigning a placard (A, B, C, etc.) to each vehicle, and require the vehicles to stay in sequence behind their predecessor, we illustrate a virtual circuit. Note that pedestrians in crosswalks, red lights, and cross traffic can interrupt and delay portions of the convoy. However, all vehicles arrive at the destination in their scheduled sequence, albeit with varying delays.

We use the terms "switching" or "routing" because the circuits, messages, and/or packets traverse a mesh of links and nodes. Links connect nodes. Nodes are also called switches. We route or switch the entities from one node to another, using the links, from the source, ultimately to the destination. The entire mass of links and nodes is a point-to-point network. Typically, wide area networks (WANs) are point-to-point networks. Another kind of network is a broadcast network. Local area networks (LANs) are typically broadcast packet networks.

In Figure 1, several nodes are shown with links connecting them. Not all nodes are directly connected to every other node. There may or may not be any loops in the network. Some nodes may be connected to only one other node. However, there are sufficient links so that, directly or indirectly, all nodes are connected to each other and form one complete network.

In Figure 2, several nodes are shown. All nodes are contained within a cloud and share a common medium or ether. A pretty, puffy cloud artistically depicts the network's edges. What is important is that each node is inside the cloud. Each node can, via the common medium, communicate directly with every other node. There is no privacy. Everyone can listen to everything.

Layering allows us to look at only a portion of the communications process at any one time. This simplifies our examination of something very complex. For example, on the data link, MAC (medium access control), and/or network layers, we may use an assortment of protocols, one on each layer or sublayer, as appropriate to the specific network. On the internet layer, we use IP (Internet protocol), a datagram protocol. On the transport layer, we use TCP (transport control protocol), a virtual circuit protocol. On the application layer, we employ SMTP (simple mail transport protocol), a message-switching technique, to ultimately transport our electronic mail messages.

Just as a real network is constructed using several different layers, each layer has its own set of addresses (e.g., hardware addresses, IP addresses, and port addresses) to specify the location of devices and services. In the case of the Internet, addresses don't always neatly map into layers.

We start by looking at the OSI and the TCP/IP models because communications networks are built in layers. For a better foundation, we also quickly look at broadcast networks, frame relay, ATM, and addressing, especially those kinds of addresses used on the Internet. Next, we look at circuit switching, briefly at message switching, and finally at packet switching in more detail.

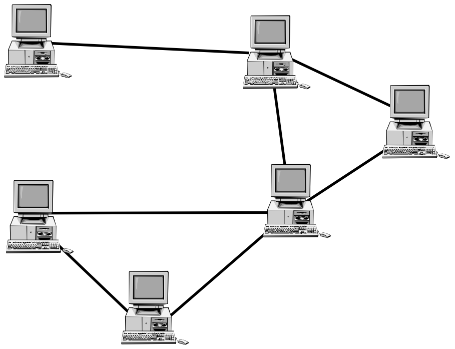


Figure 1: A point-to-point network.

CIRCUIT, MESSAGE, AND PACKET SWITCHING

Before tackling the prime topic of packet switching, it is worthwhile to discuss circuit and message switching. This provides a context for better understanding why packets were invented and remain useful.

OSI, TCP/IP, AND OTHER LAYERING MODELS

We build and discuss networks using layers, just as houses are built on foundations, usually resting on earth. A house's frame rests on the foundation, the interior and exterior walls hang on the frame, and the roof sits on top. Table 1 outlines the OSI reference model and shows a comparison to several versions of the TCP/IP model. Several TCP/IP model versions exist because it is an informal, loosely defined structure.

OSI Reference Model

The ubiquitous layering model is the seven-layer OSI reference model: physical, data link, network, transport, session, presentation, and application. This reference model, with its strictly separated and well-defined layers, was created to be a de jure blueprint for constructing point-to-point networks. However, the folks building systems that use this plan did not prevail and the TCP/IP Internet, built on its own plans, won the popularity contest. Today, the OSI reference model is used for its great pedagogic value. The realities of the world and expediency motivate us to depart from the OSI model for the construction of real networks.

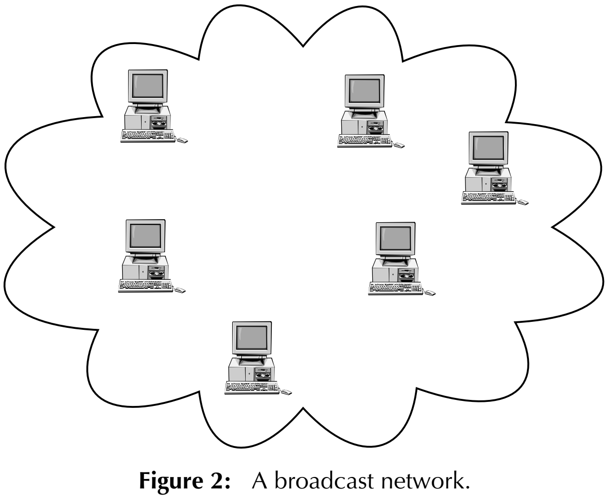


Figure 2: A broadcast network.

The OSI reference model was constructed with only point-to-point, homogeneous networks (wide area networks) in mind. It ignores the concept of a broadcast network and the idea of an internet (i.e., a network of networks constructed on top of networks). Because the OSI model is so specific and limited, we need to be especially careful when we discuss topics beyond its scope.

TCP/IP Model

The TCP/IP model is informal and is presented by different folks as a five- to three-layer model, including application, transport, and internet layers, with sometimes a data link layer and sometimes a physical layer (see Table 1). The Internet (an internet) has no network layer because it is a network, an internet, built on top of other (likely heterogenous) networks. The supporting networks might or might not have their own network layer. One supporting network could be a point-to-point WAN that likely has a network layer. Another supporting network could be a LAN, without a strictly defined network layer because it is a broadcast network (e.g., an ethernet).

The documents defining the Internet, the RFCs, standardize all sorts of things. They define techniques and protocols for the application, transport, and internet layers, and also some standards tying the internet layer to the layer supporting it (e.g., an ethernet). The definition of an ethernet is not made in the RFCs (requests for comments) but is specified elsewhere.

This chapter discusses the data link, network, internet, transport, and, briefly, application layers. The physical layer is of no interest to us here. In today's world, the presentation layer as its own entity is pretty much ignored. The session layer is almost always considered a part of the transport layer. This leaves us the data link, network, transport, and application layers from the original OSI reference model and the internet layer from the TCP/IP model to examine for circuit-, message-, and packet-switching techniques.

We examine the different layers before we get into the topics of circuit, message, and packet switching, because we most likely employ different kinds of switching on the different layers that build our network.

Broadcast Model

The OSI reference model was constructed with point-to-point (WANs) in mind. It ignores broadcast networks, which are not constructed with links joining nodes (Figure 1), but by a set of nodes all sharing the same medium (Figure 2). The OSI model can be made suitable for broadcast networks by splitting the data link layer into two sublayers, the MAC and the LLC (logical link control).

MAC sublayer protocols determine how a node or station gains access to the broadcast medium needed to transmit packets into the network. LLC sublayer protocols define a language that stations can use to decode the packets. It also provides a common layer, or glue, for

Table 1 The OSI Reference Model and Versions of the TCP/IP Model

OSI Layer	TCP/IP 5-layer	TCP/IP 4-layer A	TCP/IP 4-layer B	TCP/IP 3-layer
Application	Application	Application	Application	Application
Presentation				
Session	Transport			Transport
Transport		Transport	Transport	
	Internet	Internet	Internet	Internet
Network	Network access	Supporting networks		
Data link			Supporting networks	
Physical	Physical			

interfacing to upper layers. The LLC protocol defined by IEEE 802.2 defines a common protocol used by several different MAC schemes, such as Ethernet (IEEE 802.3), token ring (IEEE 802.5), and wireless (IEEE 802.11). There are many IEEE 802 standards. Some (e.g., 802.11) are today very rapidly evolving. A more detailed discussion is beyond the scope of this chapter.

Broadcast networks can do some things more easily than point-to-point networks can. One of these is to "broadcast" to all stations on the network. A special address designated as addressing all stations can exist. Packets with this destination address are copied by all stations in the network because all share the same medium, the same environment. In a point-to-point network, pains are taken to ensure that such broadcast packets are routed and repeated to all stations. This involves having many copies of the packet in the point-to-point network compared with needing only one in the broadcast network.

Multicasting extends the idea of broadcasting. Groups of stations are designated as sharers of a multicast address. Many such groups and many such addresses can be created. Packets sent to a multicast address are copied by all stations. Those stations recognizing the address as significant should copy the information. Other stations should ignore the message.

Packets directed to a designated destination are read by all stations. The addressed station acts on the information while other stations are expected to ignore it. Note the words "expected to ignore it." It is easy to spy and to copy everything. This is one of the motivations for using encryption on the higher layers: to keep private information private.

Frame Relay Model

Frame relay is an evolution of the X.25 networks developed using the OSI reference model. Changes in the world include the development of extensive digital switching and transmission equipment replacing older analog gear. Digital electronic switches replaced mechanical and analog electronic switches. Optical fibers replaced metallic cables. Speeds increased and errors decreased, both very significantly. The OSI reference model works by having data link frames (employing flow and error control) carrying network-layer packets (employing flow and error control), which, in turn, carry transport-/session-layer packets (employing flow and error control). Each layer has significant control overhead.

However, with digital circuits and optical cables, much faster transmission and many fewer errors exist. Why is all this repeated control work performed at each layer? Can we do it quickly enough using software computations? The frame relay solution is to collapse the data link and the network layers together and to just "switch" or "relay" the "frames" from link to link. This switching is done using hardware, not software, logic. It is performed quickly. Errors are so rare that the easiest solution to detecting an error is to discard whatever is wrong. The users (i.e., higher layer protocols such as TCP/IP) can retry when the omission is noticed.

What have we done? We've eliminated layers by merging them together. Also, we've replaced expensive, slow, software-based logic with cheaper, faster, hardware circuits.

ATM Model

If frame relay is an evolution, ATM is a revolution. In addition to the steps taken in evolving a traditional OSI-reference-model-based scheme into frame relay, we chop our data into very small, fixed-size cells (53 octets). There are two reasons for this. Hardware can manipulate the small, fixed-sized cells more easily (i.e., more quickly) than it can switch larger, varying-size frames (in the order of hundreds or thousands of octets).

Another reason for a small frame is that the transfer of CBR (constant bit rate, e.g., voice, video) traffic, in addition to bursty computer data, becomes feasible. CBR traffic is intolerant of delays that vary. Or rather it is more correct to say that it is we, human users, who are intolerant of varying delays. If a large frame of bursty computer data arrives at the switch just before a CBR frame does, the CBR data are delayed. If a small computer data cell arrives before an equally small CBR cell, there is little variability in the flow of the CBR traffic.

Table 2 lists the ATM AAL CS and the ATM AAL SAR. The AAL is the ATM adaptation layer or the application adaptation layer. It is the glue between the application and the ATM layers. The CS is the convergence sublayer. The SAR is the segmentation and reassembly sublayer

Table 2 Broadcast, Frame Relay, and ATM Models Compared to the OSI Reference Model

OSI Layer	Broadcast Model	Frame Relay	ATM
Network	DLC	Frame relay	ATM AAL CS
Data link			ATM AAL SAR
	MAC		ATM
Physical	Physical	Physical	Physical

(of the AAL layer). These two sublayers provide the glue to the layers above and below. A more complete description of ATM is beyond our scope. The ATM model changes the way data are switched on the lowest levels and provides standards to mate the ATM protocols with both CBR and computer data that utilize ATM.

TCP/IP and Other Addresses

A brief discussion of addresses gets dragged into our chapter for the purpose of clarifying where addresses fit into the scheme of packet switching. We restrict ourselves to brief peeks at Ethernet addresses, an example of hardware addresses, and IP addresses and port numbers because they are so pervasive. If we were to extend our discussion to AX.25, X.25, frame relay, ATM, or myriad other networks, we would have many other addressing schemes to examine. Even this limited discussion leaves many topics untouched.

First, let's not talk about DNS (domain naming system) addresses. An example of such a name is "gpfn.sk.ca." These mnemonic names are great for people to recognize and remember. They are not at all like the 32-bit IP addresses that they represent. IP addresses are hard to remember and to write—for human beings! IP addresses are great—for computers! The DNS is a very important scheme implemented by sophisticated Internet applications running on many computers. Additionally significant, the DNS allows us to easily change IP addresses and maintain permanence by retaining the DNS name. The computers are told about the new association and soon the change is complete.

Additionally, we will not discuss e-mail addresses, such as rhg@gpfn.sk.ca, which are composed from a character string, an @ sign, and a DNS name. Likewise, we will not discuss Web addresses (e.g., http://www.gpfn.sk.ca:80/%7Erhg/), which commonly specify a DNS name (sometimes an IP address instead), sometimes a port number, and commonly a file path designating a specific file on the computer.

Port addresses in the Internet model are 16-bit numbers that allow communications to a specific process on a host. The host is specified using an IP address (or more commonly a DNS name that is translated into an IP address). Applications bind themselves to specific ports when they want to request service in the case of client software, or when they wish to provide service in the case of server programs. (Some "well-known" ports have mnemonic names, e.g., ftp [port 21], http [port 80], smtp [port 23], on some computer systems.) The port addresses provide a glue between the Internet transport layer and application processes on the hosts.

Internet or IP addresses are 32-bit (IP version IV) numbers used to specify each host. A host can have several IP addresses, but no host can honestly exist without at least one unique IP address. IP addresses are commonly written using dotted decimal format, four decimal numbers in the range 0–255, separated with periods (e.g., 198.169.198.4). This IP address today belongs to "neale.gpfn.sk.ca," to "gpfn.sk.ca," and to "gpfn.ca." Tomorrow, the associations between the DNS names and the IP addresses may be completely different. Each device, host, computer, router, and switch on an internet has one or more IP addresses. These addresses are also used to route IP datagrams from source to destination.

00:05:02:d6:eb:7e, a 48-bit Ethernet address, written using 12 hexadecimal digits, is an example of a hardware address. This number is manufactured into each interface card. A computer has a unique Ethernet address for each interface card. Associations between IP addresses and Ethernet addresses are performed by Internet protocols that most of us, fortunately, never see. I say fortunately because usually, whenever we are interested in hardware addresses, it's because there is a problem to solve.

CIRCUIT SWITCHING

Circuit switching is easy to understand because its most common example, the voice telephone call, is so pervasive. Each call consists of three phases, establishment or setup, data transfer, the voice conversation, and termination. Establishment allocates resources to make a telephone call possible. Is the other telephone available? The called telephone could already be busy, or perhaps the line is cut or is otherwise unavailable. Is the desired person available? Are there sufficient resources within the telephone network to construct a route, or circuit, from the caller to the called station? Perhaps a natural disaster, such as a tornado or other usual weather, has caused many, many other folks to attempt phone calls, leaving the switches and trunks (the links between the switches) devoid of capacity because they are fully used by others.

The data transfer phase (i.e., the actual conversation) is the reason for making the call. The conversation usually has its own (application layer) protocol. Both folks say hello, engage in some kind of chitchat, and eventually say goodbye. The phones are replaced on their hooks. Bills may be prepared and the electronic facilities, switches, links, and so forth, are freed for other calls.

Phone calls are usually metered and billed by measuring their duration in time and sometimes the distance between the stations. The number of phonemes, words, sentences, and paragraphs transferred is not measured. Long periods of silence (i.e., no information transfer) is handled in the same way as intensive information transfer.

We can use the same voice telephone calls to transfer *computer data* (whatever that is) instead of analog voice information by using *modems*. The basic paradigm remains: establishment, data transfer, circuit termination, breakdown, and the dedication of facilities by time duration, not information transfer.

One special case needs mentioning, the "hot" or private line. Two (or sometimes more) telephones are permanently connected to each other. Facilities are always available to transfer information between the two locations. Costs for this kind of service are usually based on time (all the time) and on distance.

Why did we add the parenthetical comment "whatever that is" when talking about the transfer of computer data using "voice," analog telephone lines?

Today, telephone transmission and switching is almost universally digital, except in older facilities.

Our voice conversations are digitized, transferred, and switched as digital signals and then converted back to analog voice in time to deliver to the receiving party. Broadcast television is now (circa 2002) making the conversion to digital transmission. Broadcast radio is also slowly making this same transition.

MESSAGE SWITCHING

The great model for message switching is the paper-based postal system, which delivers cards, letters, and packages (which can contain files, books, and even electronic media). At any selected time, the information is in one location, in a carrier's vehicle or bag (moving or not), or in one of the postal stations (being sorted, waiting to be sorted, or waiting to be loaded into a vehicle for transportation). The application layer SMTP (simple mail transport protocol, e-mail) is an example of old electronic message switching that survives into the present.

Let's look at how mail delivery using SMTP works. In addition to the two folks sending and receiving the e-mail message, there are (at least) two kinds of software involved. Typically the sender employs a UA (user agent) program to compose and address the message and typically sends the message to a sender's local MTA (mail transfer agent) using SMTP. The MTA accepts the message from the UA and promises to try to deliver the message to its recipient. Usually the MTA searches for another MTA, on a distant host, willing to accept the message for the recipient. The two MTAs try to create a connection and negotiate message delivery using SMTP. This can fail for a number of reasons. The sending MTA, which still holds the message, either looks for an alternate, intermediate MTA willing to accept the message or holds onto it and tries again at a later time. The MTA holding the message is responsible for it until it is passed on to another MTA or until it returns it to the sender as being undeliverable. If the MTA passes a message on to another MTA, its responsibility for the message ends. Final delivery to the recipient's UA is done not by SMTP but by other techniques (including POP and IMAP, other e-mail protocols). This is why we added the parenthetical "at least" at the start of this paragraph.

What's important here? The message travels as a complete unit. At any one time, it exists completely in one place. Of course, we are only speaking about the application layer. What happens on the other layers is an entirely different tale!

Table 3 lists possible protocols used on each layer to transfer an e-mail message using SMTP, using TCP/IP, over a telephone connection. This same TCP/IP connection at another physical location might be supported by, say, an Ethernet LAN, because the Internet sits on top of heterogenous networks.

Table 3 Example of Different Techniques Used on Different Layers

Layer	Technique
Application	SMTP, message switching
Transport	TCP, virtual circuit
Internet	IP, datagram
Data link	PPP, virtual circuit Digitization of the analog signal within the telephone plant Modulation of an analog signal using digital data
Physical	Modem, telephone circuits, circuit switching

The application layer FTP (file transfer protocol) might be thought of as message switching because the entire file travels as a unit. However, FTP is not a good example because it is more involved with copying files than with moving them from one place to another. Also, only one FTP server is involved. In SMTP, one, two, three, or more MTAs are involved in message delivery.

The Bitnet, in the 1970s and into the 1980s, formed a worldwide network of (originally IBM) mainframes that transferred files from site to site. (An e-mail message can be regarded as a file.)

PACKET SWITCHING

We looked at circuit switching and at message switching. Why do we need to invent something new? One word: efficiency. Look at the model of the voice telephone conversation. It is always "on." Signals are always being transmitted even when both parties to the conversation are silent. What if we could reclaim those silent periods and insert other folks' conversations into the silence? With packet switching we do just that.

Think about phone conversations. There really are few silent periods. Humans insert some kind of noise (varying with the spoken human language used) to provide an indication to the other party that they are still there, still listening, and still (pretending to be) interested. Music, voice, and video information are traditionally known as CBR (constant bit rate) traffic because the smooth timing of delivery is very important to human recipients.

Think about computer online interactions, such as reading e-mail from a server using IMAP (or POP), obtaining files using FTP, or browsing the Web. A user creates a connection (without going into more detail about this) to the Internet. Using IMAP, the UA asks the mail server about received messages and gets a list. While the user stares at this list, what is happening on the computer link? Usually nothing. A particularly interesting message catches the user's eye and he asks to read it. Now the message is copied in a burst from the server to the UA and can be examined. Typically nothing happens while the recipient peruses the message, which may contain a huge

Table 4 Advantages and Disadvantages of Circuit-Switched and Packet-Switched Networks

Circuit-Switched Network	Packet-Switched Network
+ Dedicated, unshared	+ Shared
+ Constant delays	– Varying delays
+ No packet overhead	– Packet overhead

attachment. The reader decides that he wants the attachment, and then, all of a sudden, a long burst of line activity occurs to bring the attachment from the server to the client.

The previous paragraph demonstrates the bursty nature of computer communications: long periods lacking information transfer with more or less shorter or longer bursts of high-speed transfer. We want a fast link because we don't want to wait for information. However, most of the time, the (fast, expensive) line is idle because we are not using it. Packet switching was invented to cost effectively share the fast, expensive, high-speed lines with many users. In other words, packets provide something called *multiplexing*.

So what do we do? We chop our messages into units, cells, frames, packets, segments, and PDUs (protocol data units). All these words are pretty much interchangeable in general usage. Mostly our packets are of varying size (for several reasons). There are often minimum and maximum sizes (i.e., the packet is not allowed to be too small or too big). Sometimes packets are of fixed size (e.g., ATM [asynchronous transfer mode] cells are a small, fixed-size 53 octets). ATM is also called cell relay. (See Table 4.)

Datagram Networks

One thing we can do is chop our data into datagrams and spew them independently into the network. We must give each datagram two addresses: a destination address and a source address. Wherever the datagram roams in the network, the destination and the source are known because they are self-contained within the datagram. What are the advantages and the disadvantages of this scheme? On the minus side, each datagram is bigger because it contains two addresses. Also, frequently a *hop count* (or time-to-live) field is contained within the "overhead" of the datagram. The hop count is usually a positive number, which is decremented whenever a node sends the datagram onto a link toward another node. When the count is decremented to zero, the datagram is discarded. Here are two more disadvantages: Each datagram must find its own way through the network. The experience of previous datagrams (e.g., how they found a good route) is ignored. It is also possible for a datagram to wander about, to never reach its destination, and to be destroyed.

The advantages of datagrams are basically simplicity and robustness. We never perform the steps of establishing a circuit (or route) through the network. Likewise, because a circuit is never established, there is no circuit to break down and facilities to allocate. Datagrams are robust because each finds its own path through the network. If the network changes (e.g., nodes and/or links are destroyed, removed, or shut off), there is no effect on the datagrams other than it might take them longer to make the trip—as long as there is at least one possible route through the network from the source to the destination. (See Table 5.)

The IP (internet protocol) of the internet layer of the Internet employs datagrams. This makes it robust. One of its original design specifications in the late 1960s, by the U.S. Department of Defense, is nuclear survivability. Although the Internet has not been tested against a nuclear attack, destroying portions of the network, all sorts of other things have been tossed against it. It keeps on working. Recall that the internet layer is the layer that connects, sits on top of, the constituent networks. Resting on top of the internet layer is the transport layer. There are two Internet protocols on the transport layer, UDP (user datagram protocol), a datagram protocol, and TCP (transport connection protocol), a connection-oriented protocol.

Datagrams provide simplicity by avoiding the creation of a circuit. Often, we simply only want to send a small amount of information from the sender to the recipient and perhaps receive a reply. We gain by omitting the overhead of circuit establishment, maintenance, and breakdown for only a very short exchange. If we really do want a reply, when we don't get one, we just retry the whole exchange de novo.

Domain name service (DNS), the very important application that translates domain name addresses, such as gpfn.sk.ca, into IP addresses, such as 198.169,198.4, is performed using UDP/IP (i.e., datagrams on both the transport and the internet layers).

Note another characteristic of datagrams: If source A sends out a bunch of datagrams a, b, and c to destination B, there is no guarantee that any or all of the datagrams will arrive at their destination. Furthermore, we have no guarantee as to the sequence in which a, b, and/or c will arrive. This is like the Kentucky Derby. If the horses leave the starting gate at approximately the same time, we have no knowledge of the sequence in which they will arrive at the finish line. In fact, with a small probability, one or more of the horses may never cross the finish line. Sometimes this characteristic is just nice to know and is of no real consequence. At other times, it is a real concern and causes us to look toward connection-oriented services instead.

We see that datagrams have advantages and also disadvantages. There are times when it is better to use a connection-oriented protocol.

Table 5 Advantages and Disadvantages of Datagram and Virtual Circuit Networks

Datagram Network	Virtual Circuit Network
+ Efficient for short exchanges	+ Efficient for long transfers
+ Simple	– Circuit establishment and breakdown overhead
+ Robust	+ Sequenced delivery
−larger packets	+ Smaller packets

Table 6 Comparison of Datagram and Virtual Circuit Networks

Datagram Network	Virtual Circuit Network
Transfer data	Establish circuit Transfer data Breakdown circuit

Virtual Circuit Networks

TCP is the connection-oriented protocol on the Internet-transport layer, resting upon the (datagram) internet layer. Just as in circuit switching, we establish a circuit (calling it a virtual circuit), we transfer data through the circuit, back and forth, and then we disassemble the circuit and return the resources for reuse. (See Table 6.)

We do the work of allocating, maintaining, and disassembling resources with the expectation that the use we make of the virtual circuit will be sufficient to pay us back for our effort. What kind of payback can we expect? Because we have constructed a tube, like a garden hose, through the network, when the source sends segments (i.e., packets *a*, *b*, and *c*) into the pipe, we can expect them to arrive in order (i.e., *a* first, *b* second, and *c* third) at the destination. This guaranteed sequencing is an important property. These packets can also be smaller than datagrams because we no longer need the complete source and destination addresses in each packet. The virtual circuit that we initially constructed holds this information. We are saving something in the overhead in each individual packet. Each packet still needs identification, something that says which virtual circuit it belongs to and what its sequence is in the parade, in case it is destroyed in transit.

Packets in a virtual circuit scheme usually employ sequence numbers of some kind because they are part of a parade, with each packet having its own place. Packets are often numbered using three bits, 0, 1, 2, 3, 4, 5, 6; seven bits, 0, 1, 2, . . . , 125, 126; or even larger sequences, such as the 16-bit TCP sequence numbers. These sequence numbers are more overhead but are needed to identify damaged or missing packets that the receiver asks for again using a *negative* acknowledgement. The sequence numbers are also used for flow control, to moderate or slow the sender when the network and/or the receiver cannot handle the packets as quickly as they are being sent.

Virtual connections are useful in many situations (e.g., interactive telnet sessions allowing a user with a telnet client to log onto a remote computer, mail transfer using SMTP, delivery of Web content via HTTP [hypertext transfer protocol], and FTP, file transfers, are just a few examples of a small number of well-known services).

In telnet, HTTP, and SMTP, a client program opens a TCP/IP connection to a server and an application protocol (e.g., telnet, HTTP, SMTP) starts a structured conversation with the server using the “pipe.” For example, in HTTP, a browser opens a connection to a remote server and says, here is what kind of browser I am, send me this page. The server accepts the connection, parses the request, and sends the Web document (or perhaps an error) back, then closes the connection. This is the original HTTP version. However, a Web page can consist of several Web documents, graphics, frames, and/or client-side executable programs such as Java or JavaScript. The original HTTP requires a new TCP/IP connection for each document, with the overhead of establishing the connection, maintaining it, and breaking it down for each document. The newer HTTP version uses one connection and sends several documents through it, reducing the client–server overhead by using a more complicated application-layer protocol.

Permanent Virtual Circuit Networks

A permanent virtual circuit is just like a virtual circuit except that it is permanent. It is administratively constructed in the network. It is always there, whether or not it is being used. A few packets can be quickly sent down the pipe whenever needed with a minimum of fuss. The Internet does not have permanent virtual circuits. They do exist in X.25, frame relay, and ATM networks.

CONCLUSION

Let’s review what we’ve just seen.

Examples of and analogies to circuit and message switching exist in many everyday events.

Packet switching can also be illustrated with only a little more imagination.

Packet switching was invented to provide multiplexing, providing efficiency in digital networks by taking advantage of its traditionally bursty nature.

ATM was designed to accommodate CBR traffic in addition to traditional bursty computer traffic.

Packet switching can be divided into two techniques: datagrams, which provides simplicity and robustness, and virtual circuits, which offer efficiency in many common situations.

Networks, point-to-point, broadcast, and internets are naturally layered.

Each layer can be constructed using different techniques (e.g., circuit switching, datagrams, virtual circuits, and message switching).

Several different kinds of addresses are employed. Addresses may exist at particular layers or they may form a glue between layers.

Each kind of network has appropriate reference models that help us understand them.

GLOSSARY

Address Something that specifies a location or a unique instance. A person’s name, an apartment number, a house number and street name, and a telephone number are common examples.

Address, hardware An unchangeable address that is built into a device. An Ethernet address, manufactured into a network adapter, is an example. At one time, telephone numbers were hardware addresses, fixed to a specific location, albeit not to a specific telephone instrument. Today, telephone numbers are (within geographic limits) portable, not fixed.

Address, IP The Internet protocol address (version IV) is a 32-bit number assigned to a host computer on the Internet. (A host can have more than one address.) The address is used to specify a particular machine. An example is 198.169.198.4, using dotted decimal notation. Four decimal numbers, 0–255, written with dots separating them, is commonly used to write an IP address. IP addresses are used in selecting routes through the Internet from a source to a destination.

Address, port A 16-bit number. Each IP address has two complete pools of these numbers, one for TCP and one for UDP. Thus, at any one time, each IP address can sustain 4096 TCP connections and 4096 UDP transactions. The port addresses connect specific application instances, or processes, to the Internet. Each set of two IP addresses and two port addresses, at any specific time, are completely unique and specify one TCP connection.

Asynchronous transfer mode (ATM) A fast, modern networking scheme. Also called cell relay.

Bandwidth The range of frequencies, from lowest to highest, required to convey a signal. The frequency capacity of a channel. The capacity of a communications link (e.g., twisted pair, coaxial cable, optical fiber, etc.). Bandwidth is measured in units of Hertz. Increased bandwidth is strongly associated with increased data rate (measured in bits/second) and with increased signal rate (measured in Baud).

Baseband A communications link that has all of its bandwidth devoted to one communication. The link is not shared. All frequencies from the minimum (perhaps including direct current, which has a frequency of 0 Hz) to the maximum are devoted to one communication. Digital signals are often "baseband" in nature. An example is any kind of Ethernet (there are several kinds) with the notation "base" contained it its name. A more extensive definition is beyond the scope of this chapter.

Broadband A communications link that has its bandwidth shared between several communications, each having a specific frequency band. The link is shared. Frequencies from the lowest supported to the maximum are divided into bands. Analog signals can be broadband in nature. An example is television, both cable and broadcast. There is a (rarely used) version of Ethernet for cable TV with the notation "broad" contained it its name. A more extensive definition is beyond the scope of this chapter. Another, modern, use for this term is to specify fast communications. An explanation of how this usage evolved is beyond our scope.

Cell A packet, usually on one of the lower layers (e.g., data link) of the OSI reference model. Most often a cell is a small, fixed-size packet (e.g., an ATM cell).

Cell relay See ATM.

Circuit A route or path through a network. A circuit can exist in a circuit switching network or in a packet switching network. We habitually choose slightly different, and more specific, terms, such as virtual circuit or route, depending on what kind of network we are discussing.

Circuit switching The technique of establishing a route through a network that completely dedicates facilities of the links used to create the circuit entirely for one connection. The capacity of the circuit can be completely used, completely unused, or anything between. The circuit is dedicated (e.g., a telephone call).

Connected See Virtual Circuit.

Datagram An unconnected UDP packet on the transport layer of the TCP/IP model. An unconnected IP packet on the internet layer of the TCP/IP model. A generic term for a packet that travels through a network independently, from source to destination.

Frame A packet, usually on one of the lower layers (e.g., data link) of the OSI reference model. A generic term for a packet.

Frame relay A modern networking scheme.

Hardware address See Address, hardware.

Internet The worldwide network of networks employing TCP/IP. The underlying foundation networks are not homogeneous. Note the upper case "I."

internet A network of networks usually employing TCP/IP. The underlying foundation networks need not be homogeneous. Note the lower case "i."

IP Address See address, IP.

Intranet See internet.

LAN (local area network) A network in a limited geography (i.e., a room, a building, or a campus). Usually LANs are broadcast networks.

Layer Computer code, hardware, protocols (i.e., rules) that constitute an entity designed to perform a specific job. The layer is supported by lower level layers. In turn, it supports higher level layers. Ideally (OSI reference model), a layer is totally independent of its higher and lower neighbors (i.e., the layer can be entirely replaced without its peers being aware of the swap). Realistically (TCP/IP model), this is not true. TCP and UDP are useless without IP supporting it. IP has little use without TCP or UDP above it.

Layer, application The highest layer, associated with human beings and application programs.

Layer, data link The layer immediately above the physical layer, concerned with communications on links between nodes (hosts, routers, switches) of a network. This layer exists only between two nodes.

Layer, Internet The layer above supporting networks that creates a uniform internet on top of (likely heterogenous) networks. This is the IP layer.

Layer, network The layer above the data link layer. It encompasses the entire network.

Layer, physical The lowest layer, consisting of hardware and its mechanical, electrical, functional, and procedural specifications.

Layer, transport The layer above the network, or the internetwork layer, depending on context, concerned with the end-to-end delivery of packets. (UDP datagrams and TCP segments for an internet.)

MAN See metropolitan area network.

Message A complete entity. Examples include the contents of a book between its covers, the contents of an envelope in the paper-based postal system, a computer file, an e-mail (message).

Message switching The transport of a message through a network as a complete unit. The message is not broken into smaller units (i.e., packets; e.g., a letter in the postal system). The Bitnet operated as a message switching system.

Metropolitan area network (MAN) A network typically covering an area the size of a metropolis. The common example of a point-to-point MAN is the local cable TV distribution system. An example of a broadcast MAN is an FM transmitter sending information to the city. MANs are typically constructed as point-to-point networks but can be broadcast networks.

Multiplexing Various techniques for transparently sharing a transmission resource among disparate users. These include frequency and time division multiplexing for electronic signals and wave division multiplexing (WDM) for optical signals. Using packets is a form of multiplexing.

Network, broadcast A network employing a common medium or "ether." It consists of nodes and a medium accessible to all the nodes. An example is all CB radio operators tuned to a particular channel in a restricted geographic area.

Network, datagram A network organization that has packets traveling as independent units (i.e., uses datagrams). There are no circuits or connections.

Network, point-to-point A collection of nodes or switches, connected with links. The telephone system is a good example.

Network, unconnected See Network, datagram.

Network, virtual circuit A network organization that has the packets in a "connection" follow an established path through the network, in sequence, one after another.

Octet A grouping or unit of eight bits. Communications folks prefer this more specific term over the less specific "byte."

OSI reference model A seven-layer (physical, data link, network, transport, session, presentation, application) pedagogic model for networks. It is a old model that represents no modern network, but it is very useful in categorization.

Packet A "bunch" of data, usually with delimiting headers and trailers, typically tens to tens of thousands of octets, transmitted as a unit.

Packet switching The technique of breaking information into discrete "packets" and routing them through a network. The routing of the individual packets can be unconnected or connected.

Port address See address, port.

RFC, Request for Comments A constantly enlarging group of documents defining the Internet and its protocols. There are approximately 3,400 (circa August 2002) documents (http://www.ietf.org/rfc.html, retrieved August 2002).

Route See Circuit.

Routing The process of discovering and constructing a path, circuit, or route through the network from the source to the destination.

Segment A connected, TCP packet on the transport layer of the TCP/IP model.

TCP/IP model A three-layer (internet, transport, application) model for the Internet. This model sits on top of a foundation provided by other networks or LANs. Because there is no "true" standard for this model, it is variously a three-, four-, or five-layer model, which includes one or more lower layers upon which the TCP/IP layers rest.

WAN (wide area network) A network covering a large region. Usually larger than a metropolitan area. Typically it is a point-to-point network.

CROSS REFERENCES

See *Client/Server Computing; Internet Literary; Internet Navigation (Basics, Services, and Portals); Local Area Networks; Public Networks; TCP/IP Suite; Virtual Private Networks: Internet Protocol (IP) Based; Wide Area and Metropolitan Area Networks.*

FURTHER READING

Baran, Paul. (2002). The beginnings of packet switching: Some underlying concepts. *IEEE Communications Magazine, 40*(7):42–48

Beyda, W. J. (2000). *Data communications: From basics to broadband,* 3rd ed. New York: Prentice Hall.

Comer, D. E. (1997). *The Internet book,* 2nd ed. New York: Prentice Hall.

Forouzan, B. A. (2000). *TCP/IP protocol suite.* New York: McGraw-Hill.

Forouzan, B. A. (2001). *Introduction to data communications and networking,* 2nd ed. New York: McGraw-Hill.

Keogh, J. (2001). *The essential guide to networking.* New York: Prentice Hall.

Kurose, J. F., & Ross, K. W. (2001). *Computer networking.* Reading, MA: Addison Wesley Longman.

Stallings, W. (2000). *Data and computer communications,* 6th ed. New York: Prentice Hall.

Tanenbaum, A. S. (1996). *Computer networks,* 3rd ed. New York: Prentice Hall.

Click-and-Brick Electronic Commerce

Charles Steinfield, *Michigan State University*

INTRODUCTION

Despite the early fascination with dot-com companies, there is a growing recognition that the Internet is unlikely to displace traditional channels anytime soon, at least in the world of business-to-consumer (B2C) commerce. Rather, many traditional enterprises have moved to integrate e-commerce into their channel mix, using the Internet to supplement existing brick-and-mortar retail channels (Steinfield, Bouwman, & Adelaar, 2002). Electronic commerce researchers now consider the combination of physical and Web channels to be a distinct electronic commerce business model, most commonly referring to it as a "click-and-brick" or "click-and-mortar" approach (Timmer, 1998).

In this chapter, a broad overview of the click-and-brick approach to e-commerce is provided. The focus is on the use of the click-and-brick approach by firms selling consumer products and services via a combination of physical and Internet retail channels, given the relative prevalence of this situation in the e-commerce arena. Much of the discussion is also relevant to other types of companies that rely on both Internet and physical channels, such as those involved in education and health care. In the first section, a brief look at the current e-commerce situation highlights the overall importance of taking an integrated brick-and-click approach to e-commerce development. In the second section, a detailed examination of the sources of synergy between traditional and Internet-based channels is provided. The third section introduces the dangers of product and channel conflict and points out possible management strategies to improve channel integration. The fourth section highlights the potential benefits that firms may reap when pursuing a more integrated approach to e-commerce. Section five introduces four brief cases that give concrete examples of click-and-brick strategies. Section six discusses several critical factors that may inhibit firms' attempts to more tightly integrate physical and Internet sales channels. Finally, the chapter closes with several conclusions regarding the importance of the click-and-brick approach in electronic commerce research and practice.

CLICK-AND-BRICK E-COMMERCE OVERVIEW

Many types of companies can be considered click-and-brick firms, including retailers of tangible products, sellers of financial and other services, health care providers, and educational organizations extending learning services via the Internet. Even nonprofit organizations have employed a click-and-brick approach as they seek new ways of reaching and extending service to their constituents. Essentially, all click-and-brick firms have both Internet and physical outlets and seek synergies between them to reduce costs, differentiate products and services, and find new sources of revenue. In the B2C area, electronic commerce can be considered a marketing channel, which can be defined as a means to interact with end consumers. Many firms rely on a mix of different channels such as physical stores, catalog sales, and e-commerce. Firms pursuing channel integration attempt to tightly coordinate the use of channels, even within a single sales activity, to improve their profitability (Friedman & Furey, 1999). It is therefore helpful to distinguish the truly integrated click-and-brick approaches from those that treat electronic commerce more as a separate and parallel channel. The difference is illustrated in Figure 1. In the parallel case, customers are not able to move easily between electronic commerce and traditional channels. For example, many firms require that goods ordered online be returned directly to the e-commerce subsidiary, rather than through physical retail outlets.

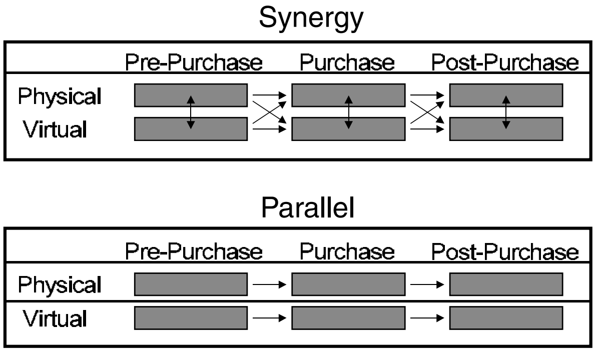


Figure 1: Contrasting synergy with parallel approaches to click-and-brick e-commerce.

In the integrated, or synergy approach, customers are able to move seamlessly between channels as they interact with a firm. For example, a customer may do product research and initiate an order online but pick up the merchandise and obtain after-sales service in a physical outlet.

In the early years of electronic commerce, many in the industry felt that pure Internet firms (the dot-coms) had significant economic advantages over traditional firms. As a result, many traditional firms chose a parallel approach to the Internet in an attempt to avoid saddling e-commerce divisions with the burdens of higher costs and reduced innovativeness that they felt characterized their traditional physical channels. The widespread failure of dot-com firms, however, forced traditional retailers to rethink this approach and seek out synergies. Laudon and Traver (2001) noted that traditional retailers are replacing all but the most established dot-coms (e.g. Amazon) on the lists of top e-commerce sites. The electronic commerce activities of traditional retailers have helped to maintain a steady growth in online sales during the period in which large numbers of dot-com enterprises have failed. The U.S. Commerce Department estimates that the number of people who have purchased a product or engaged in banking online more than doubled between 2000 and 2001, growing from 13.3% of the U.S. population in August 2000 to more than 29% in September 2001 (National Telecommunications and Information Administration [NTIA], 2002). Additionally, the NTIA (2002) reports that more than a third of Americans, and fully two thirds of the Internet users, now use the Internet to obtain product information. Not surprisingly, despite the economic slowdown in 2001 and 2002, this increased e-commerce activity has translated into growing online sales revenue. Fourth quarter 2001 e-commerce sales increased by 13.1% over fourth quarter 2000, reaching more than $10 billion. Total retail sales increased by only 5.3% during the same period. Statistics on online sales in 2002 continue to show year over year growth at a much higher rate than the overall retail sales sector (U.S. Census Bureau, 2002). (Note that this underestimates total consumer-oriented e-commerce activity, because the census does not include online travel, financial services, and ticket agencies in their retail sample.)

SOURCES OF SYNERGY BETWEEN TRADITIONAL AND E-COMMERCE CHANNELS

Click-and-mortar firms have a number of potential sources of synergy not necessarily available to pure Internet firms or traditional firms without an e-commerce channel. Borrowing from classic competitive advantage theory (see Porter, 1985), such sources of synergy include common infrastructures, common operations, common marketing, common customers, and other complementary assets that can be shared between e-commerce and physical outlets (see Figure 2).

Common Infrastructures

E-commerce channels can make use of a variety of existing infrastructures such as logistics or information technology (IT) systems to reduce costs or offer capabilities that would be difficult for dot-com firms to match. An example of the use of a common logistics infrastructure would be when a firm relies on the same warehouses and trucks for handling the distribution of goods for e-commerce activities as it does for delivery to its own retail outlets. Likewise, if a firm has a capable IT infrastructure, including product and customer databases, inventory systems, and a high-speed Internet protocol network with high bandwidth connections to the Internet, the adoption and use of Internet-based commerce can be enhanced. Firms can even use their Internet access to offer e-commerce services to customers who are actually in a store or branch, for example, via kiosks.

Common Operations

Existing retail operations can also be put to good use in support of e-commerce, permitting integrated applications to emerge. For example, an order processing system shared between e-commerce and physical channels may enable improved tracking of customers' movements between channels, in addition to potential cost savings.

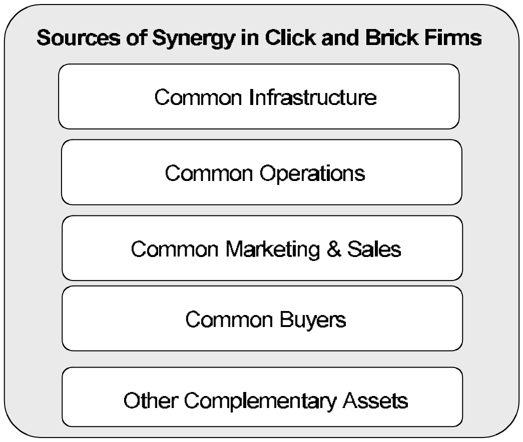


Figure 2: Sources of synergy in an integrated click-and-brick approach. Adapted from Steinfield, Adelaar, & Lai, 2002.

Common Marketing

E-commerce and physical channels may also share common marketing and sales assets, such as a common product catalog, a sales force that understands the products and customer needs and directs potential buyers to each channel, or advertisements and promotions that draw attention to both channels.

Common Buyers

Instead of competing with each other, or pursuing different target markets, e-commerce channels and physical outlets in click-and-mortar firms often target the same potential buyers. This enables a click-and-mortar firm to be able to meet customers' needs for both convenience and immediacy, enhancing customer service and improving retention.

Other Complementary Assets

There are many other types of complementary assets that click and mortar firms possess that purely Internet firms may not. The management literature, for example, notes such additional complementary assets as existing supplier and distributor relationships and experience in the market. As with the other sources of synergy, to the extent firms are able to share these assets across channels, they will be better able them to take advantage of an innovation such as e-commerce (Afuah & Tucci, 2001; Teece, 1986).

Of course, click-and-brick firms also obtain many of the same benefits from the use of the Internet channel as other Internet companies. Certainly the Internet channel affords the possibility for 24-hours-a-day, seven-days-a-week customer access; lower cost access to many new markets; the opportunity to develop and maintain a community of customers; and an efficient communication channel for customer input.

MANAGING CHANNEL CONFLICT IN MULTICHANNEL FIRMS

The integration of e-commerce with existing physical channels is a challenging undertaking that can create problems for management. More specifically, firms with multiple channels may fall prey to channel conflict. Channel conflicts can occur when the alternative means of reaching customers (e.g., a Web-based store) implicitly or explicitly competes with or bypasses existing physical channels and are nothing new to e-commerce (Stern, El-Ansary, & Coughlan, 1996). One danger is that these conflicts result in one channel simply cannibalizing sales from the other. This is particularly a problem when there are clear cost, convenience, or other advantages for customers using e-commerce, causing them to substitute Internet channels for traditional ones. Such cannibalization of sales becomes a strong threat to company viability if it is difficult to capture equivalent revenues online. Media companies such as newspaper companies, although not exactly brick-and-mortar firms in a traditional sense, have experienced such product-channel conflicts. This is exacerbated by customers' unwillingness to pay for online information or entertainment, necessitating free online versions that steal subscriptions from physically distributed media products. Some of the loss is made up by increased advertising revenue from online channels, but it has become increasingly clear that such business models are not sustainable in the current environment (Laudon & Traver, 2001). Additional, and potentially even more damaging, problems for the content industries that sell music and other nonperishable content (i.e., not news) arise from piracy and the massive and unlawful distribution of copyrighted content among Internet users. Clearly, the product-channel conflicts experienced by companies in the content industry represent a significant challenge. The review in this chapter focuses on a narrower definition of click-and-brick firms, however, in which actual click-and-mortar outlets exist to support transactions with customers. Hence, I do not deal directly with issues of copyright and piracy.

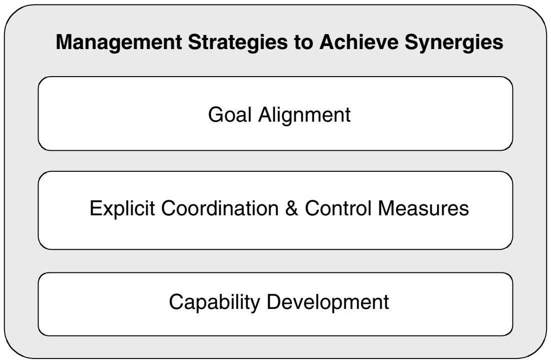

Management Strategies to Achieve Synergies
Goal Alignment
Explicit Coordination & Control Measures
Capability Development

Figure 3: Categories of management strategies used to avoid channel conflicts and achieve synergy benefits. Adapted from Steinfield, Adelaar, & Lai, 2002.

Perceived threats caused by competition and conflict across channels can have other harmful effects, including limited cooperation across the channels, confusion when customers attempt to engage in transactions using the two uncoordinated channels, and even sabotage of one channel by the other (Friedman & Furey, 1999). Management must act to diffuse conflicts and ensure the necessary alignment of goals, coordination and control, and development of capabilities to achieve synergy benefits (Figure 3; Steinfield, Bouwman, & Adelaar, 2002).

Goal Alignment

One of the first tasks for managers of click-and-brick firms is to ensure that all employees agree that an e-commerce channel is needed and that they will support it. This represents a process of goal alignment. Aligning goals across physical and virtual channels implies that all employees involved realize that the parent firm benefits from sales originating in either channel. One problem faced by click-and-brick firms is that the contributions made by the Internet channel may be intangible and difficult to measure. Managers have to be open to such intangible benefits and not, for example, evaluate e-commerce divisions purely on the basis of online-only sales and profitability. For example, the improved communication with customers offered by Internet channels may enable firms to better understand customer satisfaction and needs, yielding better products and services and long-run customer retention. Such benefits may not show up in pure Internet sales figures, however. Moreover, there must be agreement as to what types of customers (e.g., existing vs. new) are to be targeted by the new e-commerce channel. In essence, to avoid channel conflict, management must be proactive in obtaining the support of all employees, building consensus about the goals, methods of evaluating success, and targets for e-commerce. If management simply puts an e-commerce division into place without this goal alignment step, existing employees may feel threatened and may be uncooperative.

Coordination and Control Measures

In addition to obtaining consensus on goals, explicit coordination and control mechanisms are needed to move a click-and-brick firm more in the direction of integrated, rather than parallel channels. First, it is important for click-and-brick firms to design for interoperability across channels, so that customers may move freely between online and physical retail outlets. For example, customers may want to search a particular store's inventory from their own home computer to see if a specific item is in stock or not. They may want the store to hold the item for them to pick up on their way home, rather than have it delivered. This implies that the online system connects and interoperates with the store system.

Another example of explicit coordination is the use of each channel to promote the other. For example, an online visitor may be informed about various in-store special sales, just as an in-store customer may be told about particular complementary services found on the click-and-brick firm's Web site. The use of kiosks in stores, for example, can directly enable customers on premises to access additional information, services, or promotions at precisely the moment they are considering a specific purchase. Cross-promotions enhance the perception that all the arms of the click-and-brick business are working together to add value for customers. They also create payoffs from e-commerce that can reduce resistance, such as when online promotions generate greater in-store traffic.

One of the most critical management issues for click-and-brick firms is to provide real incentives to employees to encourage cross-channel cooperation. Imagine a situation where store managers know that any time a customer buys something online instead of in the store, it represents lost revenue and lower compensation. Store personnel will invariably encourage customers to buy in the store, because this provides employees with real income. E-commerce sales may help improve customer relations, but no rational salesperson will knowingly direct customers to a sales channel that only ends up reducing his or her income. In many successful click-and-brick firms, efforts have been made to allocate online sales to particular establishments, so that e-commerce does not succeed at the expense of physical outlets. This often is possible when customers have accounts tied to a particular establishment, and online sales from specific accounts are credited to the home establishment. Another way of allocating online sales to a physical outlet is to use the address of the customer.

Finally, click-and-brick firms often find that they are able to capitalize on the unique strengths of each sales channel, affording the possibility for some degree of channel specialization (Steinfield, Bouwman, & Adelaar, 2002). For example, costs for certain types of transactions may indeed be less in an online environment, suggesting that companies should encourage customers to use more efficient channels when possible. Banks, for example, have long attempted to persuade customers to use ATMs for routine cash transactions, rather than coming to a branch and occupying the services of a teller. On the other hand, many financial transactions require expert advice and counseling and are best done in-person at a branch. Hence, click-and-brick banks following a channel specialization strategy might offer customers incentives such as better interest rates or lower fees in return for the use of more efficient online channels for routine transactions like money transfers or bill payments. When more "advice-sensitive" transactions are sought, customers would be directed to their local physical bank branch (Steinfield, Bouwman, & Adelaar, 2002).

When there is potential for severe product-channel conflicts, such as when customers decide to forego a newspaper or magazine subscription to read free content online, a channel specialization strategy may offer a viable alternative. For example, some newspaper and magazine companies derive additional revenue from their online versions using the unique search and archive advantages of the Web-based channel. A specific niche segment of their market may willingly pay for the ability to research back issues and articles on certain topics. This approach to adding value to an online version to encourage payment has been tried by such publications as the *Wall Street Journal* (Laudon & Traver, 2001).

Capability Development

In many situations, traditional firms may lack important competencies needed to achieve synergy benefits with e-commerce. For example, traditional firms may lack Web development skills or logistics skills needed to serve distant markets. Indeed, the ability to serve remote customers is an important prerequisite to successful Web-based commerce, which explains why catalog companies such as Lands' End have experienced so much success with e-commerce. It fits well with their existing capabilities and adds value in the form of enhanced transaction efficiencies and lower costs. When lacking such capabilities, alliances may be more useful than attempting to develop a virtual channel in-house. Managers must recognize whether the requisite competencies are present in the existing traditional company and, if a partner is needed, must carefully construct an alliance that ensures that their e-commerce partner is not simply siphoning business from physical retail outlets.

POTENTIAL BENEFITS OF AN INTEGRATED CHANNEL APPROACH

Once click-and-brick companies recognize the various sources of synergy across channels and develop management strategies to avoid conflicts and encourage

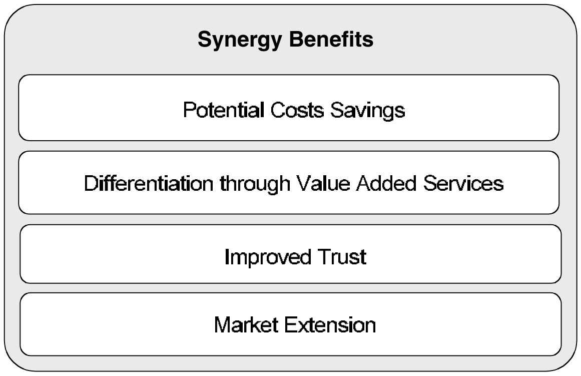


Figure 4: Categories of potential benefits from an integrated click-and-brick approach. Adapted from Steinfield, Adelaar, & Lai, 2002.

cooperation across channels, numerous benefits may result. Four broad areas of benefit include (a) lower costs, (b) increased differentiation through value-added services, (c) improved trust, and (d) geographic and product market extension. The potential benefits from physical and virtual integration are depicted in Figure 4 and discussed in this section.

Lower Costs

Cost savings may occur in a number of areas, including labor, inventory, marketing and promotion, and distribution. Labor savings result when costs are switched to consumers for such activities as looking up product information, filling out forms, and relying on online technical assistance for after-sales service. Inventory savings arise when firms find that they can avoid having to stock infrequently purchased goods at local outlets, while still offering the full range of choices to consumers via the Internet. Marketing and promotion efficiencies are garnered when each channel is used to inform consumers about services and products available in the other. Delivery savings may result from using the physical outlet as the pickup location for online purchases or as the initiation point for local deliveries.

Differentiation Through Value-Added Services

Physical and virtual channel synergies can be exploited at various stages in a transaction to help differentiate products and add value. Examples of prepurchase services include various online information aids to help assess needs and select appropriate targets, or, conversely, opportunities in the physical environment to test products. Examples of purchase services include ordering, customization, and reservation services, as well as easy access to complementary products and services. Postpurchase services include online account management, social community support, loyalty programs, and various after-sales activities that may be provided either online or in the physical store. Typical opportunities are in the areas of installation, repair, service reminders, and training. Although many of these value-added services are potentially available to single-channel vendors, combined deployment of such services (e.g., online purchase of computer with in-store repair or training) can enhance differentiation and lock-in effects (Shapiro & Varian, 1999).

Across all transaction stages, the improved communications from the Internet channel (assuming a well-developed and fully functional site) can yield better insights into customer needs and wants. Such customer relationship management can ultimately add value for both parties over the lifetime of a customer's interaction with a click-and-brick firm, offering benefits in excess of the actual online sales by enhancing loyalty and retention.

Improved Trust

Three reasons for improved trust, relative to pure Internet firms, derive from the physical presence of click and mortar firms, including reduced consumer risk, affiliation with and embeddedness in recognized local social and business networks, and the ability to leverage brand awareness. Lower perceived risk results from the fact that there is an accessible location to which goods can be returned or complaints can be registered. Affiliation and embeddedness in a variety of social networks may facilitate the substitution of social and reputational governance for expensive contracts or legal fees (Granovetter, 1985). For example, an online customer may be more prone to trust the Web site of a business when the store manager is a member of his or her church or when the customer's business and the online vendor are both members of the same chamber of commerce. Such ties are more likely to exist between geographically proximate buyers and sellers, suggesting that there may indeed be a preference for doing business with firms that are already physically present in the local market. Finally, marketing theorists have long recognized the power of branding as a means of building consumer confidence and trust in a product (Kotler, 1999). Established firms are able to leverage their familiar name to make it easier for consumers to find and trust their affiliated online services.

Geographic and Product Market Extension

Adding a virtual channel can help extend the reach of a firm beyond its traditional physical outlets, addressing new geographic markets, new product markets, and new types of buyers. Those in other geographic markets may be new or former customers who have moved away. Virtual channels can also extend the product scope and product depth of physical channels by enabling firms to offer new products that they do not have to stock locally. Moreover, firms may add new revenue-generating information services online that would not be feasible to offer in physical outlets. Finally, the Internet may help reach customers within an existing market who may not have visited the physical outlet but are otherwise attracted to the virtual channel due to its special characteristics.

SUMMARY OF THE CLICK-AND-BRICK FRAMEWORK

The click-and-brick framework elements can be assembled into the summary framework portrayed in Figure 5.

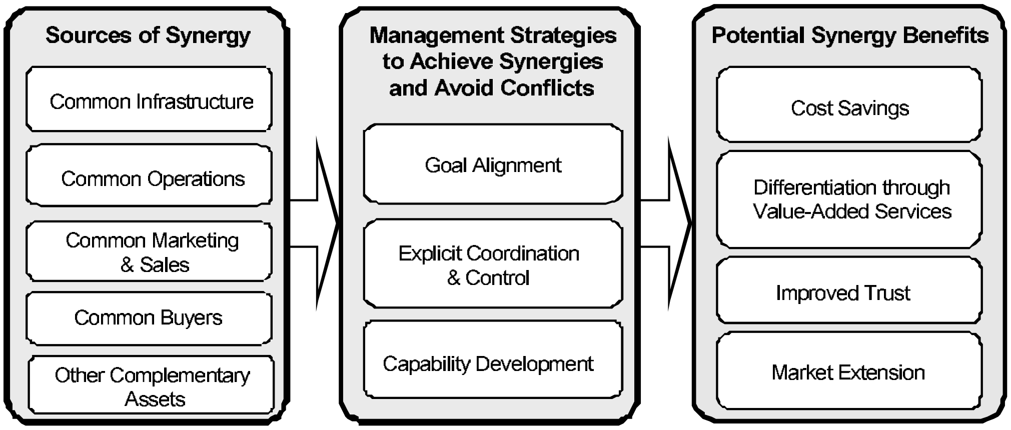

Sources of Synergy
Common Infrastructure
Common Operations
Common Marketing & Sales
Common Buyers
Other Complementary Assets
Management Strategies to Achieve Synergies and Avoid Conflicts
Goal Alignment
Explicit Coordination & Control
Capability Development
Potential Synergy Benefits
Cost Savings
Differentiation through Value-Added Services
Improved Trust
Market Extension

Figure 5: A comprehensive click-and-brick model. Adapted from Steinfield, Adelaar, & Lai, 2002.

The framework directs our attention to the sources of synergy, the need for management strategies to capitalize on click-and-brick applications, and the potential benefits that can result.

EXPLORING THE FRAMEWORK WITH SEVERAL CLICK-AND-BRICK CASES

The framework is best illustrated by describing several cases of click-and-brick firms. These examples were selected and adapted from a series of click-and-brick cases developed by Steinfield, Adelaar, and Lai (2002). Among their cases (the actual firm names were not reported based on the wishes of the interviewed companies) were a specialty retailer, a business-to-business (B2B) building materials supplier, an automobile manufacturer selling through dealerships, and a financial services firm. Each represents a different basic business arrangement—a multichannel retailer operating in the B2C arena, a B2B wholesaler, a manufacturer promoting sales to consumers through an affiliated, but independently owned, network of dealerships, and a firm selling information and services rather than tangible products. They nicely illustrate the robustness of the click-and-brick approach.

An Electronics Retailer

This company is one of the largest specialty retailers of consumer electronics, personal computers, entertainment software and appliances with more than 400 stores. The firm recently rolled out a new e-commerce site that featured both a deeper selection of products and a tighter integration with its traditional physical stores. The click and mortar design strategy enables the firm to benefit from a range of synergies between its virtual and physical channels. The goal is to be “channel agnostic,” letting customers choose whichever channel or combination of channels best suits their needs.

A number of sources of synergy are available to the firm. One key source is the firm’s exploitation of a common IT infrastructure between its e-commerce and store channels. It accomplishes this by tightly integrating the Internet operations with existing databases and other legacy systems. The firm also consciously capitalizes on common operations, especially in terms of purchasing, inventory management, and order processing. That common marketing and common buyers were a source of synergy is evident in the emphasis on replicating and leveraging the store brand in its online services.

One of the services enabled by the tight IT integration allows online customers to check out the inventory of individual stores, so that they might order merchandise for immediate pickup in the nearest store. To achieve this value-added service and derive the differentiation benefit from it, the service had to be supported by a change in business processes that ensured interoperability across the two channels. For example, if only one or two items that an online purchaser desires are in stock, in-store customers might claim them by the time the Web customer arrives for pickup. To avoid this situation, store personnel must be notified that an online customer has requested an item for pickup. Then employees remove the item from the shelf and send an e-mail confirmation to the online customer. To ensure that stores cooperated with this new capability, management incentives were also considered to avoid or diffuse potential channel conflicts. In particular, the company included performance in fulfilling online orders as one of the parameters influencing store manager compensation.

This seemingly simple service thus reflects the main components of the framework. Several sources of synergy come into play. First, the firm built the service by tying the Internet to a common, integrated IT infrastructure. Second, it supports the service by using existing store inventory that was warehoused and delivered using common logistics infrastructure. Third, the shoppers can provide payment that is credited to the store, using existing operational systems such as credit card verification and approval systems already in place. Finally, the service targets common buyers, that is, people living near existing physical stores.

Management initiatives to achieve synergy and avoid conflict are also evident in this simple example. The online service depended on the cooperation of store personnel, reflecting a need for goal alignment. This was achieved by developing a service that brought traffic into the store, rather than simply bypassing it altogether. Moreover, management recognized that the Internet could assist in pre-purchase activities, even if the eventual sale was consummated in the store. It did not require the e-commerce channel to generate its own profits. Additionally, it attended to the need for explicit coordination and control by developing a business process that ensured cross-channel interoperability. Finally, it created an incentive system that rewarded store personnel for their cooperation with the e-commerce channel.

The benefits of this one service are also captured well by the framework. Consider the cost savings in labor that stores accrue when customers search for products online, conduct research, order the product, and even make payment ahead of time, all without needing the assistance of a single employee. In terms of differentiation, this is represents a prepurchase and purchase service that would be difficult for a non–click-and-mortar firm to offer. It caters to the different needs and wants of store and online customers, improving satisfaction. Because of the tie-in to the local store, which is also part of a well-known national chain, customers perceive much lower risk than they would if ordering from a less familiar, nonlocal Internet business.

The tight integration between the e-commerce and existing retail infrastructure offers this firm many other advantages that are derived from the same sources of synergy and enabled by many of the same management strategies. For example, because of the integrated approach, customers who order products online with home delivery are able to return products to their local store, enhancing trust and reducing perceived risk. Moreover, the integration of IT systems enabled store employees to access customer and order data to improve customer assistance, such as finding complementary goods.

Channel cooperation extends in both directions. In-store customers who are unable to find a product on the shelf can search the firm's online site through kiosks available in the store. Because of the integrated approach to marketing, the firm is also able to undertake promotional campaigns, such as sales and contests that customers can access in the store and on the Web. In addition, the Web channel also enables value added services geared toward improving customer relationship management. In particular, the Web site allowes customers to store items under consideration in a "Think About" folder. This provides useful marketing information to the firm, because it can provide more targeted promotions related to desired products.

A Building Material Supplier

The building materials supply company has a double-pronged approach to e-commerce. First, it maintain its own Web site, offering rich information services to its primary customer base—the professional builder. Second, it has an alliance with a building supply portal that allows it to offer e-commerce transactions to its existing client base as well as to new customer segments.

Professional builders are provided with an account that allows them to use the e-commerce site. They can log in directly from the builder supply firm's home page. A local lumberyard where the builder has an account fulfills the orders. Essentially, each lumberyard caters to the market located within a radius of 100 miles. Prices are individualized, encouraging builders to consolidate their purchases for volume discounts.

The building supply portal works with other suppliers but is tightly coupled with the case study firm because both firms have the same principle stockholder. In addition to providing online supply ordering services to builders, it also enables the firm to offer value-added services, extend into the consumer home improvement market, and provide customers goods that the local lumberyard does not carry. Among the value-added services are a variety of accounting and management options that builders can use. These include maintaining an online ledger for a project—for example, a house—that can be used as a template for the next project, saving builders time on order entry. Each builder can have his or her own personal Web page, including a personal product usage and construction plan folder. Builders can also check the status of their orders on a daily basis and order material outside the regular store hours. This is helpful because many builders do their administrative tasks at home in the evening. The personal pages include information on activities and promotions occurring at the local branch.

By outsourcing consumer e-commerce transactions to the online portal, the company now has a presence in the growing home-improvement market dominated by such superstores as Home Depot. Because of the other partners who participate in the portal, the company is able to offer their existing customers one-stop shopping services, even for goods that they do not carry.

Individual stores receive electronically all orders placed on the portal for their products and fulfill them through their normal supply chain and existing fleet of delivery trucks. New professional clients are first encouraged to set up an account at a local branch, where local employees negotiate individual pricing arrangements.

The Web-based service supports standard orders. The company still maintains outside sales representatives (OSRs), however, who visit with builders on job sites to maintain good customer relations. The increased use of online ordering by builders allows OSRs to pay less attention to administrative tasks and to focus on selling value-added services, giving advice, educating the client about the online channels, and strengthening customer relationships.

Through its online partner, the firm also completed a successful mobile service pilot using Palm Pilots. Builders were able to make on-site purchases for critical materials needed immediately. Materials were then brought out to the job by the local delivery truck. One interesting impact of this service is that it encouraged builders to wait until the last minute for some orders and to make orders in smaller quantities than they would through normal channels. This is, of course, less efficient for the supply firm, creating some challenges for their delivery system. It is, however, a new value-added service that strengthens their relationship with core clients.

An Automobile Manufacturer

All automobile manufacturers realize that car shoppers are able to conduct extensive research online before buying a car. Carmakers have well-developed Web sites providing rich information about their models, but for a variety of reasons they are not able to sell cars directly to end customers through the Web. Hence, they must work with traditional dealers to offer a click-and-brick experience. In one carmaker's e-commerce service, customers can configure their desired car online, obtain a fixed price quote, and choose a local dealer from whom they wish to take delivery. The application locates the matching car from dealer inventories, and if in stock at a different dealer from the one chosen by the customer, the dealers will swap cars with each other. The chosen dealer then gets full credit for the sale of the online-configured car, as well as the continuing service relationship to the customer. At the Web site, customers can also research cars and check the inventory of local dealers online. In addition, customers can apply online for credit and insurance, which is also submitted to local dealer.

This approach helps the manufacturer sell more cars without alienating its existing dealer network. The company realizes that in the car market, because of the logistics of delivery and the need for a physical presence for service and warranty work, bypassing dealers will not work. In fact, in many states, it is illegal for car manufacturers to sell directly to end consumers. To secure support from dealers for the initiative, an e-dealer advisory board was created. The manufacturer has also introduced features such as online scheduling for maintenance and repair and an ownership Web site where customers can find accessories that go with their car and receive maintenance service reminders.

Other electronic services initiatives include the introduction of mobile in-car services for safety, security, and information. They are combining the Global Positioning System with wireless technology to deliver emergency roadside assistance, stolen vehicle tracking, navigation aids, and other travel-related services. The selling dealer will activate these in-car wireless services and provide training to customers.

A Financial Services Provider

The final case described here is a large national bank offering a traditional range of banking and financial services. The company focuses on the consumer and small business segment and has more than 8 million households as customers. It took a somewhat different path from the other cases to arrive at a click-and-brick strategy. In fact, this bank, through acquisitions and mergers, had both a typical online banking service available to its account holders and an entirely separate Internet brand that solely targeted new customers over the Web. In the description of this case that follows, the former is referred to as the internal online banking service, whereas the latter is referred to as the Internet pureplay bank.

Their Internet pureplay bank was not successful, largely because of the high costs of customer acquisition and the lack of necessary synergies with the parent bank. One particularly troublesome problem faced by all Internet-only banks results from the difficulties and costs of transferring money into and out of the Internet bank. Because of customers' reluctance to deposit cash or checks in ATMs, Internet-only bank users first have to deposit funds in traditional banking account and then transfer money (e.g., via a check or bank transfer) to their Internet account. This implies that the Internet pureplay bank would always be a supplementary rather than a primary bank. Moreover, in this bank's earlier parallel approach, the Internet pureplay bank had to rely on other banks' ATM networks to allow customers to make cash withdrawals, resulting in relatively high surcharges.

On the other hand, the internal online banking channel service was closely tied to the brick-and-mortar branches, had much lower customer acquisition costs, and offered many customer retention services. Existing account holders were freely offered the opportunity to sign up for online access to their bank accounts and other banking services, allowing online customer acquisition at a fraction of the cost of finding entirely new bank account clients. Customer retention was enhanced by providing such services as online bill payment. Services such as this increase retention by raising switching costs, because customers need to reenter extensive billing information if they change to a new bank. Another synergy benefit the bank experienced with its internal online banking channel stemmed from cost savings at brick and mortar branches due to the ability to offload routine transactions to the cheaper and more convenient Web channel. In keeping with the channel specialization management strategy described earlier, price incentives were also introduced to stimulate the use the online services.

The parent bank took advantage of IT infrastructure synergies between its branches and its integrated online banking service by designing financial applications once and implementing them across the virtual and physical channels. For example, an application to speed up the approval of home equity loans was "mirrored" between the Web and the physical branches. The bank effectively developed coordination and control measures to be able to offer seamless customer support, so that customers would not have to repeat any transactions on multiple channels. For example, if customers changed their address using one channel, all systems would be updated at the same time. To achieve this, the bank had to integrate systems so that Web services were integrated into the day-to-day operations of branches.

As a result of these experiences, the formerly separate Internet pureplay bank was reintegrated into the parent bank, tying it more closely to the click-and-mortar branches. Networked kiosks in bank branch offices introduced the Internet brand more directly to existing bank clients. The parent bank now supplies many of the core services, including deposits and withdrawals, and its connection to the Internet brand creates the trust that had been lacking. The benefit for customers is that they have the choice of a more differentiated set of banking services, with more differentiated prices that reflect the costs of transactions.

CONCLUSION

This chapter has provided a broad overview of the click-and-brick business approach, noting both the potential benefits as well as the challenges that firms face because of channel conflicts. I introduced a framework to help understand the dynamics of managing a click-and-brick enterprise. The framework begins by identifying potential sources of synergy available to firms that choose to integrate e-commerce with traditional forms of business. It further emphasizes the many actions that firms can take to minimize channel conflicts and help achieve the benefits of synergy, and it describes four categories of synergy-related benefits from the integration of e-commerce with traditional businesses, including potential cost savings, gains due to enhanced differentiation, improved trust, and potential extensions into new markets. Four case studies were described to provide a concrete illustration of the approaches taken by click-and-brick firms.

ACKNOWLEDGMENT

I am grateful to the reviewers and editor for the thoughtful comments on an earlier version of this chapter.

GLOSSARY

Channel A means by which a seller interacts with end consumers. Many firms rely on a mix of different channels, such as physical stores, catalog sales, and e-commerce. Firms pursuing channel integration attempt to coordinate the use of channels tightly, even within a single sales activity, to improve their profitability.

Channel conflict Channel conflicts occur when an alternative means of reaching customers (e.g., a Web-based store) implicitly or explicitly competes with or bypasses existing physical channels. Perceived threats caused by competition and conflict across channels can have other harmful effects, including limited cooperation across the channels, confusion when customers attempt to engage in transactions using the two uncoordinated channels, and even sabotage of one channel by the other.

Channel specialization Click-and-brick firms that attempt to direct customers to the most appropriate channel (e.g., one that is the lowest cost or one that offers the requisite capabilities) are pursuing a channel specialization approach. It allows firms to capitalize on the unique strengths of each sales channel.

Complementary assets Assets possessed by a firm, such as existing supplier and distributor relationships and experience in the market, that help it take advantage of innovations such as e-commerce.

Differentiation A competitive approach used by companies to set themselves apart from competitors through higher quality products and better customer services. Click-and-brick firms hope that they can use their combined channels to differentiate themselves from competitors.

Kiosks Self-service computer stations, often with a touch-screen display, that are located in malls, stores, and other places where customers can use them to locate products or information and access services electronically. Click-and-brick firms may offer networked kiosks on store premises to allow in-store customers access to the firms' e-commerce channel.

Synergy When the combined effect of two actions is greater than the sum of the individual effects. Click-and-brick firms hope that by combining traditional and online services, they can offer an experience to customers that is greater than possible through each channel by itself.

CROSS REFERENCES

See *Business-to-Business (B2B) Electronic Commerce; Business-to-Business (B2B) Internet Business Models; Business-to-Consumer (B2C) Internet Business Models; Collaborative Commerce (C-commerce); Electronic Commerce and Electronic Business; Electronic Data Interchange; Electronic Payment; E-marketplaces.*

REFERENCES

Afuah, A., & Tucci, C. (2001). *Internet business models and strategies: Text and cases*. New York: McGraw-Hill Irwin.

Friedman, L. G., & Furey, T. R. (1999). *The channel advantage: Going to market with multiple sales channels to reach more customers, sell more products, make more profit*. Boston: Butterworth Heinemann.

Granovetter, M. (1985). Economic action and social structure: The problem of embeddedness. *American Journal of Sociology, 91,* 481–510.

Kotler, P. (1999). *Marketing management* (10th ed.). Upper Saddle River, NJ: Prentice Hall.

Laudon, K., & Traver, C. (2001). *E-commerce: Business, technology, society*. Boston: Addison-Wesley.

National Telecommunications and Information Administration (2002). *A nation online: How Americans are expanding their use of the Internet*. Washington, DC: Author. Retrieved November 1, 2002, from http://www.ntia.doc.gov/ntiahome/dn

Porter, M. E. (1985). *Competitive advantage: Creating and sustaining superior performance*. New York: Free Press.

Shapiro, C., & Varian, H. (1999). *Information rules: A strategic guide to the network economy*. Boston: Harvard Business School Press.

Steinfield, C., Adelaar, T., & Lai, Y. (2002, January). Integrating brick and mortar locations with e-commerce: Understanding synergy opportunities. *Proceedings of the Hawaii International Conference on Systems Sciences* [CD-ROM]. Washington, DC: IEEE Computer Society Press.

Steinfield, C., Bouwman, H., & Adelaar, T. (2002). The dynamics of click and mortar e-commerce: Opportunities and management strategies. *International Journal of Electronic Commerce, 7,* 93–119.

Stern, L. W., El-Ansary, A. I., & Coughlan, A. T. (1996). *Marketing channels*. Upper Saddle River, NJ: Prentice Hall-International.

Teece, D. J. (1986). Profiting from technological innovation: Implications for integration, collaboration, licensing and public policy. *Research Policy, 15,* 285–306.

Timmer, P. (1998). Business models for electronic markets. *Electronic Markets, 8*(2), 3–8. Retrieved November 1, 2002, from http://www.electronicmarkets.org/modules/pub/view.php/electronicmarkets-183 (registration required).

U.S. Census Bureau (2002). *Service sector statistics*. Retrieved November 1, 2002, from http://www.census.gov/mrts/www/current.html

Client/Server Computing

Daniel J. McFarland, *Rowan University*

INTRODUCTION

The name of an information system often describes the utility it provides. The functionality of a transaction processing system, a decision support system, and an executive information system are self-evident. However, client/server computing is broadly defined; rather than describing system utility, client/server describes the system's architectural configuration. As a result, a client/server system may incorporate a broad range of technologies and address a variety of business situations.

Client/server computing is a form of cooperative processing. Specifically, a client/server system includes at least two software processes working together to provide application functionality. At the most basic level, a client software process requests services from a server software process. In turn, the server process supplies services to the client process. The service request may provide access to an organizational resource, such as a database, a printer, or e-mail.

Client/server computing involves the coordination of assorted computing/networking devices and the coordination of various software processes. The client/server *system* architecture describes the physical configuration of the computing and networking devices and the client/server *software* architecture describes the partitioning of an application into software processes.

The system architecture focuses on the physical architecture of the system. In this context, clients and servers are seen as computers rather than software processes. Simple client/server system architectures include a client (e.g., a personal computer) that links to a server (e.g., a mid-range computer) using a network (e.g., a local area network). Each computer provides processing and memory resources to the client/server application. The system architectural perspective views a client/server application as a network of computers sharing resources to solve a problem.

The client/server *software* architecture describes the partitioning of an application into software processes. In this context, clients and servers are seen as software processes rather than computers. These processes may reside on the same computer or they may be distributed across a network of computers. The software architectural perspective views a client/server application as a set of programming modules sharing resources to solve a problem. This chapter focuses on the client/server *software* architectural perspective; consequently, a client shall refer to a software process that requests services from other software processes and a server shall refer to a software process that provides services to other software processes.

This chapter begins with a review of three common client/server classification methodologies. The first methodology classifies a client/server application based on the computational role of the constituent processes. In particular, how an application distributes its presentation, application processing, and data services determines how the application is classified. The second methodology classifies an application based on the number of constituent processes; this is known as the tier-based classification. Finally, a server-based classification methodology is presented; it describes a client/server application in terms of the server process utility.

The subsequent section describes three broad categories of enabling technologies. The first set of enabling technologies, middleware, provides a client/server application the ability to integrate client processes with server processes. The second set of enabling technologies, component software, provides self-contained, self-describing software modules that may be combined to form applications. The third set of enabling technologies, networking technologies, bridges the gap among the computing and networking devices.

The final part of this chapter reviews intranets, extranets, and the Internet, three popular client/server implementations.

CLIENT/SERVER CLASSIFICATION

A client/server classification provides information about the particular system architectural configuration. Three client/server classifications are reviewed below. The first classification employs the presentation–application–data (P–A–D) architecture, which describes the computational role of the client process. The second classification describes the number of separate processes in a client/server application, otherwise known as the number of tiers.

Table 1 P–A–D Architectural Classification of Client/Server Applications

Computational Service	Distributed Presentation	Local Presentation	Distributed Application Logic	Local Application Logic	Distributed Data
Data services	Server	Server	Server	Server	*Shared*
Application services	Server	Server	*Shared*	Client	Client
Presentation services	*Shared*	Client	Client	Client	Client

The final classification describes the functional role of the server process.

Presentation–Application–Data (P–A–D) Architecture

Client/server computing divides an application into two or more separate processes. The P–A–D architecture describes the computational role of both the client and server processes. In particular, the P–A–D architecture defines the computational requirements of an application in terms of presentation services, application services, and data services. The presentation services layer details the interaction between the user and the application. The application services layer includes the embedded programming and business logic. The data services layer describes the connectivity to organizational resources such as data, printers, and e-mail (Goldman, Rawles, & Mariga, 1999).

Presentation Services

Presentation services represent the interactions an application has with the external environment. At the most basic level, these interactions involve user input and application output. User input represents the portion of the user's knowledge submitted to an application. Input provides application boundaries, operating parameters, and/or data. Application output strives to provide the user/process/application with new knowledge. Output includes printed reports, audio signals, screen displays, and data. In many cases, a client/server application receives input from and provides output to multiple sources, such as users, other software processes, and/or other applications.

Presentation services can also influence the degree to which a person will use an application. A complex, synergistic relationship exists among the user, the business problem/context, and the application (Dix, Finlay, Abowd, & Beale, 1998). Several areas of study explore how people and technology interact, such as cybernetics, human–computer interaction, ergonomics, human factors analysis, technology acceptance, and industrial engineering.

Application Services

Application services describe how input is transformed into output. The application services layer represents the procedural knowledge (i.e., business logic) embedded in the application. These services include statistical, mathematical, graphical, and/or data transformations.

Data Services

Data services provide access to organizational resources including information-based resources (e.g., databases), hardware-based resources (e.g., printers), and communication-based resources (e.g., e-mail). The data services layer provides utility to the application services layer and/or the presentation services layer.

Organizational resources are often difficult to access as a result of security, budgetary, and/or technical constraints. Security rules restrict access to sensitive data such as payroll information. Budgetary constraints restrict accessibility to consumable resources such as printers. Additionally, technology-based resources, such as databases, often require special and/or proprietary interfaces. As a result, data services often attempt to hide the implementation difficulties associated with accessing and using organizational resources.

P–A–D Architecture-Based Client/Server Classification

The P–A–D architecture classifies client/server applications based on the distribution of computational services between the client process and the server process. As seen in Table 1, the P–A–D architectural framework includes five client/server classification levels. Each level is defined by the client-side P–A–D functionality.

In the first level, *distributed presentation,* the client shares the presentation processing with the server and the server handles all application and data services. In the second level, *local presentation,* the client handles the presentation services and the server handles the application and data services. In the third level, *distributed application logic,* the client addresses the presentation services and a portion of the application services; the server handles the data services and shares the application services with the client. In the fourth level, *local application logic,* the client handles the presentation and application services and the server handles the data services. In the last level, *distributed data services,* the client handles all presentation and application services and the data services are shared by the client and the server.

Tier-Based Client/Server Classification

Client/server computing partitions an application into two or more separate processes. The tier-based classification methodology views each software process as a computational component/layer/tier within the overall application.

A 2-tier system is the simplest and oldest client/server architecture. In a 2-tier system a client communicates with a server directly. This architecture often strives to provide reusable, robust links to organizational resources, such as printers and data.

In a 3-tier system, a client communicates to a server through an intermediary process. This intermediary process provides functionality as well as connectivity.

Middle tier functionality may include message queuing, application execution, and database staging. The middle tier also provides encapsulation for the data source and abstraction for the presentation services. As a result, changes to the data structure and/or user interface are less disruptive and more transparent. Furthermore, data encapsulation improves connectivity with legacy and/or mainframe systems. A 3-tier system may support a large number of concurrent users and is able to handle very complex processing (Carnegie Mellon Software Engineering Institute, 1997d).

The *n*-tier client/server architecture is a natural extension of the 3-tier architecture. An *n*-tier system has an undetermined number of intermediary processes and server processes. A client process might call upon a middle tier process, which in turn might call upon another middle tier process. In combining narrowly focused processes together, the *n*-tier architecture strives to provide application flexibility and software reusability.

Server Functionality Client/Server Classification

Many client/server applications strive to reuse server process functionality. Since a server process often embeds and hides complicated implementation details, reusing a process significantly reduces system development time and effort. Furthermore, reusing successful, robust server processes improves overall system reliability (Weiss, 2001).

To promote reusability and clarity, the functional role of a server process often describes the server. The following list describes a few server types. Table 2 provides a summary of the servers listed below.

Web Server

A Web server creates a standardized way for objects to communicate with each other over the Internet. A Web server provides session connectivity to clients running Web browsers. Early Web servers fetched files for clients without processing or interpreting the content. However, technologies such as common gateway interface (CGI), Active Server Pages (ASP), JavaServer Pages (JSP), Java Applets, ColdFusion, extensible markup language (XML), and Web services have dramatically expanded the processing and interactivity capabilities of Web servers.

Application Server

Application servers provide interactivity and processing functionality to clients. An application server may execute programming logic, store and process information, and/or manage user interactivity. Centrally storing programming logic allows developers to reuse existing, proven application functionality as well as simplifying software version control and software deployment.

Wireless Server

A wireless server provides network and Internet connectivity for wireless devices. Two popular wireless servers are *wireless enablers* and *wireless gateways*. A *wireless enabling* server stores information specifically designed for wireless devices, such as wireless markup language (WML). WML is a wireless-friendly version of hypertext markup language (HTML). A wireless client accessing a wireless enabling server would be provided with a WML file rather than an HTML file. Unlike a wireless enabler, a *wireless gateway* typically does not store content. It accepts requests from wireless and traditional devices and routes each request to an appropriate server based upon the request and device type. Some wireless gateway servers also translate wireless and traditionally formatted messages to/from WML and HTML. These translations allow wireless devices to access both traditional and wireless content.

Transaction Processing (TP) Monitoring Server

A transaction processing (TP) monitoring server introduces performance efficiencies when serving a large number of users. These servers accept client requests, merge

Table 2 Functional Descriptions for Several Server Types

Server Type	Functional Description
Web server	Connectivity and interactivity to clients running Web browsers
Application server	Interactivity and processing functionality
Wireless server	Wireless content locating services and/or wireless-friendly content
Transaction processing (TP) server	User responsiveness, processing prioritization, scheduling, and balancing
Message server	Message routing and delivery
E-mail server	Electronic-mail storage, routing, and delivery
Fax server	Fax storage, routing, and delivery
Proxy server	Intermediary content and processing functionality
Firewall	Access restriction based on a predetermined set of security criteria
Dynamic host configuration protocol (DHCP) server	Dynamic reuse of network addresses
File transfer protocol (FTP) server	File transfer capabilities

requests together by type, then transmit the resulting consolidated messages to appropriate servers. In addition, a TP monitor might monitor, prioritize, schedule, process, balance, and/or rollback service requests.

Message Server

A message server accepts and routes messages. A message is a service request that is packaged with destination information. An e-mail request is a message; it has embedded intelligence regarding the specific message destination. As a result, an e-mail server is a message server.

E-mail Server

An e-mail server accepts, routes, stores, and delivers electronic mail (e-mail) messages; it handles incoming and outgoing e-mail messages. Several e-mail standards exist such as messaging application programming interface (MAPI), X.400, and the Internet e-mail standard. Any compliant software product, regardless of the authoring product, may read standardized messages.

Fax Server

A fax server is similar to an e-mail server. It accepts, routes, stores, and delivers faxes; it handles incoming and outgoing faxes. Most fax servers provide additional features such as a direct-to-print option (i.e., allowing a fax to be sent to a printer rather than an electronic inbox).

Proxy Server

A proxy server is an intermediary process. It intercepts client requests and attempts to satisfy each request itself. If a request cannot be satisfied, the proxy server routes the request to an appropriate server.

A proxy server may control and manage Internet usage. Since all client requests go through the proxy server, an organization is able to monitor and/or limit Internet usage. A proxy server that filters Internet requests and/or logs Internet usage is acting as a firewall.

Firewall

A firewall reduces the risks associated with unauthorized access to a network and inappropriate use of networking resources. All requests entering or leaving a network must pass through a firewall. The firewall rejects those requests failing to satisfy the predetermined security criteria. A firewall might also record request and network activities. Firewalls are highly configurable based upon organizational security and auditing concerns.

Dynamic Host Configuration Protocol (DHCP) Server

Like a postal address for a brick-and-mortar building, each computer on a network must possess a unique address. Traditionally, these addresses were fixed, meaning (1) each computer required individual configuration, (2) potential conflicts were inevitable, and (3) organizations were required to manage large blocks of addresses.

A dynamic host configuration protocol (DHCP) server dynamically assigns Internet addresses to computers on the network. This permits an organization to standardize computer configurations, to reduce network address contention, and to continually recycle a smaller number of network addresses.

File Transfer Protocol (FTP) Server

A file transfer protocol (FTP) server transfers files from one computer to another using a network. An FTP server accepts file uploads (i.e., copying files to a server) and file downloads (i.e., copying files from a server). While many FTP servers provide unrestricted access, others provide access to authorized users only.

ENABLING TECHNOLOGIES

Enabling technologies allow heterogeneous software processes, computing devices, and networking devices to interoperate. Enabling technologies include technological standards, organizing frameworks, and product implementations. Three enabling technologies are discussed below: middleware, component software, and networking technologies.

Middleware

Middleware is a primary enabler of client/server computing. Middleware acts like glue holding together the separate software processes. Middleware products simplify the design, development, and implementation of complex applications by hiding server location issues, networking protocol details, and operating system differences (Carnegie Mellon Software Engineering Institute, 1997a). Middleware is broadly defined; it may be a type of program-to-program interface, a set of standards, or a product. Table 3 provides a summary of the middleware categories listed below.

Transactional Middleware

Transactional middleware supports transaction-processing functionality; it strives to provide responsiveness and data integrity while supporting a large number of users. It often includes load-balancing, replication, and/or two-phase commits. The Customer Information Control System (CICS) from IBM is a transactional middleware product. CICS is common on IBM mainframes and on several other IBM-platforms including OS/2, AS/400, and RS/6000. The distribute transaction processing (DTP) protocol from the Open Group is another transactional middleware standard. Many relational and object-oriented database management systems support DTP. BEA Tuxedo, from BEA Systems, is another a popular transactional middleware product.

Table 3 Middleware Categories and Implementation Examples

Middleware Category	Implementation Examples
Transactional middleware	CICS—IBM DTP—Open Group BEA Tuxedo—BEA Systems
Procedural middleware	ONC RPC—IETF DCE RPC—Open Group
Message-oriented middleware	WebSphere MQ—IBM JMQ—Sun Microsystems

Message-Oriented Middleware

Message-oriented middleware (MOM) supports client/server processes using asynchronous, peer-to-peer messages. A MOM does not require continuous, active communication between the client and the server. A client sends a message and does not wait for a response. If a server is busy or unavailable, received messages wait in a queue until they are satisfied. MOM solutions are well suited for event-driven and object-oriented applications. However, many MOM implementations are proprietary. As a result, they tend to be inflexible, difficult to maintain, and they lack interoperability and portability (Carnegie Mellon Software Engineering Institute, 1997b). MOM middleware products include WebSphere MQ (formerly MQSeries) from IBM and Java Message Queue (JMQ) from Sun Microsystems. WebSphere MQ is a mature product supporting over 35 platforms (http://www.ibm.com). JMQ is based on Java Message Service (JMS), which is a Java application programming interface (API) included in the Java 2 Platform, Enterprise Edition (J2EE). JMS is one of the many standards included in J2EE, which is a platform for developing distributed applications using the Java programming language. J2EE, using the CORBA/IIOP communications infrastructure, specifies many services including asynchronous communications using JMS, naming services using Java Naming, Directory Interface (JNDI), transaction services using Java Transaction API (JTA), and database access services using Java Database Connectivity (JDBC). J2EE implementations are developed and provided by a variety of vendors (Sun, J2EE, n.d.; Alur, Crupi, & Malks, 2001).

Procedural Middleware

A remote procedure call (RPC) supports synchronous, call/wait process-to-process communications. When using an RPC, a client requests a service then waits for a response. As a result, an RPC requires continuous, active participation from the client and the server. Each RPC request is a synchronous interaction between exactly one client and one server (Carnegie Mellon Software Engineering Institute, 1997c). Open Network Computing Remote Procedure Call (ONC RPC), developed by Sun Microsystems now supported by the Internet Engineering Task Force (IETF), was one of the first RPC protocols and is widely deployed to a variety of platforms. The Distributed Computing Environment (DCE), developed and maintained by the Open Group (formerly known as the Open Systems Foundation), is a set of integrated services supporting the development of distributed computing environments. DCE provides distributed services and data-sharing services including RPC, directory services (i.e., the ability to identify and locate system resources), time services, security services, and thread services (i.e., the ability to build concurrent applications). DCE data-sharing services include diskless support and distributed file system services (Open Group, 1996). DCE RPC is the protocol used by several object and component middleware solutions, such as Microsoft's Distributed COM (DCOM) and COM+ and it is also an optional protocol in the Object Management Group's Common Object Request Broker Architecture (CORBA).

Component Software

Component software strives to improve software development productivity through the reuse of self-describing, self-contained software modules called components. Each component provides specific functionality. An application calls upon components at run-time.

Component software exists within a component model, which defines the way components are constructed and the environment in which a component will run. The component model consists of an interface definition language, a component environment, services, utilities, and specifications. An interface definition language (IDL) defines how a component interacts with other components to form applications. The components exist within a component environment. A component environment is a computing environment providing the ability to locate and communicate with components. Popular component models include Microsoft's Component Object Model (COM), the Object Management Group's CORBA, Sun Microsystems's Enterprise Java Beans (EJB), and Web services.

Component Object Model (COM)

Microsoft's COM is a way for multivendor software components to communicate with each other. It provides a software architecture that allows applications to be built from binary components supplied by different software vendors. COM provides the foundation for many technologies including object linking and embedding (OLE), ActiveX, and Microsoft transaction server (MTS). Distributed COM extends COM by allowing remote component interactions. COM+ further expands the COM by encapsulating DCOM, MTS, and a host of other services including a dynamic load-balancing service, an in-memory database, a publish-and-subscribe events service, and a queued-components service. COM components require the COM application server, available on the Windows operating system platforms (Microsoft, 1998).

Common Object Request Broker Architecture (CORBA)

CORBA is an open, vendor-independent architecture describing object-oriented process interactions over a network. CORBA, supported by the Object Management Group (OMG), utilizes a standard communication protocol IIOP (Internet inter-ORB protocol). IIOP allows a CORBA process running on one computer to interact with a CORBA process running on another computer without regard to network, vendor, programming language, or platform heterogeneity. CORBA supports a variety of services including asynchronous and synchronous, stateless and persistent, and flat and nested transactions (OMG, n.d.).

Enterprise Java Beans (EJB)

Similar to CORBA, Sun Microsystems's EJB is a specification. An EJB must be programmed using Java and must be hosted on a J2EE application server. EJBs interact with each other using Java remote method invocation (RMI). An RMI is a set of APIs used to create distributed applications using Java; it is also an implementation that operates in a homogeneous environment using Java Virtual Machine (JVM). RMI allows Java objects to communicate remotely with other Java objects. Alternatively, an EJB built with EJB 2.0 specification, or higher, may communicate using the message-oriented-middleware, JMS (Marinescu & Roman, 2002).

An EJB system consists of three parts, the EJB component, the EJB container, and the EJB object. An EJB component exists within an EJB container running on an EJB server. The EJB object provides access to the EJB components. An EJB component is a software module that provides specific functionality; it may be discovered and manipulated dynamically at run-time. The EJB container provides the execution environment for the EJB components; it handles the communication and interface details. An EJB container may support more than one EJB component. While an EJB component runs on a server, an EJB object runs on a client. An EJB object remotely controls the EJB component (Sun, RMI).

Web Services

Web services are an extremely promising set of technologies; they may represent the first successful cross-platform component architecture. A Web service is a self-contained program providing a specific function that is packaged in such a way that it may be called upon remotely. A Web service is a platform-independent module that utilizes existing Internet standards, primarily XML, to provide, validate, and interpret data; hypertext transfer protocol (HTTP) to provide synchronous communications; and simple mail transfer protocol (SMTP) to provide asynchronous communications.

Web services provide Internet-accessible, building block functionality to developers; they allow for incremental development and deployment. Web services support new application development and promise to provide access to existing applications. By wrapping traditional processes in Web service protocols, an existing computing infrastructure can improve interoperability and utilization without having to reprogram existing applications. Furthermore, an existing application may enhance functionality by integrating with new Web services.

Web services are distinctive in several ways. While a traditional server process is designed to provide function to a particular client, a Web service strives to provide global access to a very specific function. A Web service is not duplicated or moved; it is simply called upon as required at run-time by applications. Furthermore, unlike other distributed application environments, such as CORBA, J2EE, and COM+, Web services do not dictate the underlying communications framework (Singh, Stearns, Johnson, et al., 2002).

Open, standardized interfaces are the heart of Web services. Standards define how a Web service communicates, how it exchanges data, and how it is located. Simple object access protocol (SOAP) is an XML-based protocol for exchanging information in a distributed environment. SOAP is platform independent; it consists of an envelope, a set of encoding rules, and a convention describing procedure calls. A SOAP envelope describes what's inside the message and how the message should be processed. The encoding rules describe the application-specific data types. The conventions describe remote procedure calls and responses. Other standards include the Web services description language (WSDL), which defines the Web service interfaces and the universal description, discovery and integration (UDDI), which allows developers to list/post new Web services and to locate existing Web services (Ewald, 2002).

Table 4 Internet Networking Categories and Components

Internet Networking Category	Components
Network access	Internet access providers Internet connectivity: Dial-up, ISDN, ADSL, HFC, wireless, T1/T3
Network core	Routers and bridges Data routing: circuit-switching, packet-switching
Network edge	Computers, users, applications Internet protocol stack

Networking

The Internet consists of host computers, client computers, routers, applications, protocols, and other hardware and software. As an organizing framework, these technologies may be categorized using three broadly defined functional areas, namely network access, network core, and network edge.

As seen in Table 4, network access describes how devices connect to the Internet. The network core describes the intercommunications among the computing and networking devices. The network edge describes those objects using the Internet such as clients, servers, and applications (Kurose & Ross, 2001; Peterson & Davie, 1999).

Network Access

An Internet access provider (IAP) provides Internet access. The structure of an IAP is roughly hierarchical. At the lowest level, a local Internet service provider (ISP) provides connectivity for residential customers and small businesses. The local ISP connects to a regional ISP. The regional ISP connects to a national/international backbone provider (NBP). The various NBPs interconnect using private connections and/or public network access points (NAP) to form the backbone of the Internet.

To communicate with an IAP, a connection must be established and maintained. Several connection alternatives exist including dial-up, integrated services digital network (ISDN), asymmetric digital subscriber line (ADSL), hybrid fiber coax (HFC, via a cable modem), radio wireless, cellular wireless, satellite wireless, microwave wireless, T1, and T3.

The Network Core

The network core describes how networking devices interconnect with each other. The backbone of the Internet is an extremely large mesh. Data are sent through this mesh using either circuit-switching or packet-switching channels. A circuit-switching network establishes and

Table 5 Comparison of the OSI Model and the Internet Protocol Stack

OSI Model	Internet Protocol Stack	Layer Description	Protocols
Application	Application	Specifies how applications and processes will communicate with each other	HTTP, SMTP, MIME, DNS, NFS. Middleware products: RPC, MOM, CORBA/IIOP, CICS, DTP
Presentation			
Session			
Transport	Transport	Specifies how information is transmitted	TCP, UDP
Network	Network	Specifies host and application addressing	IP
Link	Link	Specifies how the media are shared	PPP, SLIP, ATM, Ethernet
Physical	Physical	Specifies the physical characteristics of the media	100BaseT, IEEE802.5, IEEE802.11, Token Passing, TCS, PDM

maintains a circuit (i.e., a specific network path) for the entire duration of an interaction. A circuit-switching network knows where to look for messages and the data are received in the order in which they are sent. A packet-switching network divides each transmission into a series of individually addressed packets. Since each packet is addressed, the network path need not be determined and packets may be sent using whatever bandwidth exists at the time the message is transmitted. As a result, each packet might take a different network path. The receiver of the message is responsible for unpackaging and sequencing the packets.

The Network Edge

The network edge consists of those objects using the Internet such as clients, servers, and applications. To communicate, objects must adhere to the same networking protocol. A protocol is an agreed-upon set of rules governing a data transmission. The Internet supports a variety of protocols. Some protocols are competing, meaning they serve the same purpose; others are complementary, meaning they serve different purposes.

Internet Protocol Stack

A protocol stack is a framework used to organize and describe networking protocols. Each layer in a protocol stack serves a particular networking need. Furthermore, the lower layers of the stack provide a foundation for the upper layers (Kurose & Ross, 2001; Peterson & Davie, 1999). As seen in Table 5, the Internet protocol stack is a variation of the widely accepted OSI model. The OSI model has the following seven layers: physical, link, network, transport, session, presentation, and application. The Internet protocol stack has the following five layers: physical, link, network, transport, and application. The application layer on the Internet protocol stack envelops the application, presentation, and session layers of the OSI model. A description of each layer of the Internet protocol stack follows.

Physical Layer. The physical layer of the Internet protocol stack describes the physical characteristics of the networking media. Media can be broadly categorized as guided or unguided. Guided media includes copper and fiber optics. Unguided media includes radio, microwave, and satellite. The media type influences bandwidth, security, interference, expandability, contention, and costs. Examples of physical layer protocols include 10BaseT, 100BaseT, Gbit Ethernet, IEEE802.5 Token Passing, IEEE802.11 Wireless, transmission convergence sublayer (TCS) and physical medium dependent (PMD) for asynchronous transfer mode (ATM), and transmission frame structures T1/T3.

Link Layer. The link layer of the Internet protocol stack describes how devices share the physical media. These protocols define host–host connections, host–router connections, and router–router connections. The primary role of link layer protocols is to define flow control (i.e., how to share the media), error detection, and error correction. Link layer protocols include point-to-point (PPP) and serial line internet protocol (SLIP), which allow clients to establish Internet connections with Web servers. Ethernet and ATM are also popular link layer protocols.

Network Layer. The network layer of the Internet protocol stack defines host and application addressing. The Internet protocol (IP) assigns a unique number, called the IP address, to devices connected to the Internet. The IP address uniquely identifies every computer connected to the Internet.

Transport Layer. The transport layer of the Internet protocol stack defines how information is shared among applications. The Internet transport protocols include transmission control protocol (TCP) and user datagram protocol (UDP). TCP is a connection-oriented protocol. A TCP-based application establishes and maintains a connection for the entire duration of each transaction. This connection provides reliability and congestion control. Applications using TCP include e-mail, file transfers, and Web browsing. UDP is a connectionless protocol. An UDP-based application communicates by sending individual datagram packets. Datagrams support applications requiring large bandwidth and those tolerating some degree of data loss. Applications using UDP include telephony, streaming video, and interactive games.

Application Layer. The application layer of the Internet protocol stack defines application–application communications. For example, the HTTP allows Web clients to locate, retrieve, and interpret HTML objects and other Web-based objects (e.g., VBscript, Java Applets, JPG files). SMTP sends and stores Internet e-mail messages. Multipurpose Internet mail extension (MIME) supports e-mail attachments. Post office protocol (POP3) and Internet mail access protocol (IMAP) allow clients to download e-mail messages from e-mail servers. FTP allows file transfers between computers. The domain name system (DNS) resolves IP addresses and universal resource locator (URL) names. Telnet allows for remote terminal access over the Internet. The network file system (NFS) protocol allows for remote storage of files.

CLIENT/SERVER IMPLEMENTATIONS

Internet

The Internet is a network of networks. Hundreds of millions of computers, users, and applications utilize the Internet. The Internet utilizes a variety of protocols including TCP/IP, HTTP, and FTP. The Internet supports Web browsing, online chatting, multiuser gaming, e-mail delivery, fax delivery, and telephony.

Intranet

An intranet is a private network using the same technologies as the Internet. Since the foundation, protocols, and toolsets are the same, an intranet looks, feels, and behaves like the Internet. The difference between an intranet and the Internet is scope. An intranet user is limited to visiting only those Web sites owned and maintained by the company hosting the intranet, whereas an Internet user may visit any publicly accessible Web site. Since intranets are private, individuals outside the organization are unable to access intranet resources. As a result, organizations may publish internal and/or confidential information such as organizational directories, company policies, promotion procedures, and benefits information. A firewall protects an intranet from the nonsecure, public Internet.

Extranet

An extranet is an intranet that provides limited connectivity with the Internet. Specifically, an extranet allows authorized Internet users to access a portion of the company's intranet. This access is granted through a firewall. An Internet visitor submits a username and password to the firewall. If authenticated, the visitor is permitted limited access to the company's internal intranet. Similarly, an extranet may provide authorized internal users limited access to the Internet.

Extranets often support field personnel and employees working from home. In addition, extranets are designed to interconnect suppliers, customers, and other business partners. The exposure of the extranet is based on the relationships a company has with its partners and the sensitivity of the content stored on the extranet. In the most relaxed scenario, individuals gain access to common business materials by completing a brief survey. In more restrictive scenarios, an organization may sign a contract detailing acceptable use of the extranet or consumers may be required to pay subscription fees.

CONCLUSION

Client/server computing is so broadly defined, that it says little about a particular system. A client/server application may provide national security services to a government or it may provide publicly available editorials and commentaries. Additionally, client/sever computing encompasses several architectural interpretations. The client/server *system* architecture describes the physical configuration of the computing and networking devices. In this context, clients and servers are seen as computers. The client/server *software* architecture describes the configuration of the software processes. In this context, clients and servers are seen as software processes. Furthermore, client/server computing utilizes an enormous array of heterogeneous technologies, protocols, platforms, and development environments. While some client/server applications are relatively simple, involving only a few technologies, others are extremely complex, involving a large variety of technologies.

Common to all client/server systems is the partitioning of an application into separate processes. Two general categories of processes exist, those that request services (i.e., clients) and those that provide services (i.e., servers). A client process works together with a server process to provide application functionality.

A client/server classification provides information that describes a particular system. A classification may describe the computational role of the processes, the physical partitioning of an application into tiers, or the functional role of the processes. While these classification methodologies help describe systems, they also highlight the variety and complexity among client/server implementations. The computational and functional role of the processes varies from system to system. Furthermore, a client may interconnect with many servers and a server may interconnect with many clients.

Enabling technologies strive to provide seamless integration of hardware and software in complex environments. Specifically, middleware governs the interactions among the constituent software processes, component software provides the ability to create applications using existing, self-contained software modules, and networking technologies bridge the gap among the computing and networking devices.

Early distributed computing consisted of vendor- and platform-specific implementations. However, the trend in client/server computing is to provide standardized connectivity solutions. Web services represent one of the most promising new client/server developments. Web services strive to utilize the communication infrastructure of the Internet to deploy client/server applications. While it is too early to know whether Web services will fulfill their potential, they promise to change the way application programs are developed and deployed. As the excitement and investments in such technologies continue to grow, the standardization and globalization of client/server computing will likely become a reality in the near future.

GLOSSARY

2-tier An application that is partitioned into two separate processes that work together to provide system functionality.

3-tier An application that is partitioned into three separate processes that work together to provide system functionality.

Client process A software program that requests services from one or more server processes.

Client system A computer that supports a client process and requests services from one or more server systems.

Client/server software architecture The logical and physical partitioning of an application into software processes.

Client/server system architecture The architectural configuration of computing and networking devices supporting a client/server application.

Client/server tier A well-defined, separate process representing a portion of a client/server application.

Component A self-contained, self-describing software module that provides specific functionality, where several can be combined to build larger applications.

Component environment A computing environment providing the ability to locate and communicate with components.

Component model The way components are constructed and the environment in which a component will run, consisting of an interface definition language, a component environment, services, utilities, and specifications.

Enabling technology Provide a client/server application with the ability to locate, communicate with, and/or integrate the constituent processes.

Extranet An intranet that provides limited access to/from the Internet using firewall technology.

Interface definition language (IDL) Defines how a program may communicate with a component.

Internet protocol stack A subset of the OSI model, and a framework that organizes and describes Internet networking protocols.

Internet A global network of networks supporting hundreds of millions of computers, networking devices, applications, and users.

Intranet A private network based on Internet protocols and technologies.

Middleware The software governing the interaction among the client and server processes.

***n*-tier system** An application that is partitioned into an unspecified number of separate processes that work together to provide system functionality.

OSI model A general framework that organizes and describes networking protocols.

P–A–D architecture Defines an application in terms of the following three computational services: presentation services, application services, and data services.

Process An executing program.

Server process A software program that provides services to one or more client processes.

Server system A computer that supports a server process and provides services to one or more client systems.

CROSS REFERENCES

See *Databases on the Web; Extensible Markup Language (XML); Extranets; HTML/XHTML (HyperText Markup Language/Extensible HyperText Markup Language); Intranet; Local Area Networks; Middleware; TCP/IP Suite; Wireless Internet.*

REFERENCES

Alur, D., Crupi, J., & Malks, D. (2001). *Core J2EE patterns*. Upper Saddle River, NJ: Prentice-Hall.

Box, D., Ehnebuske, D., Kakivaya, G., et al. (2000, May 8). *Simple object access protocol (SOAP) 1.1*. Retrieved August 1, 2002, from http://www.w3.org/TR/SOAP/

Carnegie Mellon Software Engineering Institute (1997a). *Software technology review: Middleware*. Retrieved March 14, 2002, from http://www.sei.cmu.edu/str/ descriptions/middleware_body.html

Carnegie Mellon Software Engineering Institute (1997b). *Software technology review: Message-oriented middleware*. Retrieved March 25, 2002, from http://www.sei.cmu.edu/str/descriptions/momt_body.html

Carnegie Mellon Software Engineering Institute (1997c). *Software technology review: Remote procedure call*. Retrieved March 25, 2002, from http://www.sei.cmu.edu/str/descriptions/rpc_body.html

Carnegie Mellon Software Engineering Institute (1997d). *Client/server software architectures–An overview*. Retrieved March 25, 2002, from http://www.sei.cmu.edu/str/descriptions/clientserver_body.html

Dix, A., Finlay, J., Abowd, G., & Beale, R. (1998). *Human-computer interaction(2nd ed.)*. New York: Prentice Hall.

Ewald, T. (2002). *Understanding XML Web services: The Web services idea, Microsoft Corporation*. Retrieved March 22, 2003, from http://msdn.microsoft.com/webservices/understanding/readme/default.aspx

Goldman, J. E., Rawles, P. T., & Mariga, J. R. (1999). *Client/server information systems: A business-oriented approach*. New York: Wiley.

Kurose, J. F., & Ross, K. W. (2001). *Computer networking: A top-down approach featuring the Internet*. New York: Addison Wesley.

Marinescu, F., & Roman, E. (2002). *EJB design patterns: Advanced patterns, processes, and idioms*. New York: Wiley.

Microsoft COM Technologies (1998). *DCOM*. Retrieved February 12, 2002, from http://www.microsoft.com/com/tech/dcom.asp

OMG (n.d.). CORBA. Retrieved February 12, 2002, from http://www.corba.org/

Open Group (1996). *What is distributed computing and DCE?* Retrieved March 3, 2002, from http://www.osf.org/dce/

Peterson, L. L., & Davie, B. S. (1999). *Computer networks: A systems approach* (2nd ed.). San Francisco: Morgan Kaufmann.

Singh, I., Stearns, B., Johnson, M., et al. (2002). *Designing Enterprise applications with the J2EETM platform* (2nd ed.). Boston: Addison Wesley.

Sun (n.d.a). *Java 2 Platform, Enterprise Edition (J2EE)*. Retrieved August 1, 2002, from http://java.sun.com/j2ee/

Sun (n.d.c). *Java remote method invocation (RMI)*. Retrieved August 1, 2002, from http://java.sun.com/j2se/1.4.1/docs/guide/rmi/

Weiss, A. (2001, January 29). *The truth about servers*. Retrieved April 25, 2002, from http://www.serverwatch.com/tutorials/article.php/1354991

Collaborative Commerce (C-commerce)

Rodney J. Heisterberg, *Notre Dame de Namur University and Rod Heisterberg Associates*

COLLABORATIVE COMMERCE TRENDS FOR THE REAL-TIME ENTERPRISE

Collaborative commerce (c-commerce) is a strategy for the next stage of electronic business (e-business) evolution. C-commerce business practices enable trading partners to create, manage, and use data in a shared environment to design, build, and support products throughout their life cycles, working separately to leverage their core competencies together in a value chain that forms a virtual enterprise. Sustainable competitive advantage may be realized by adoption of c-commerce strategies and business models.

Today forward-looking enterprises are seeking to achieve the benefits of c-commerce by leveraging the Internet as an enabling technology in order to transform their core business management decision making practices. These real-time enterprise value propositions are based on managing information rather than inventory. The business case for c-commerce is predicated on information technology (IT) investments to facilitate collaborative business practices with shared demand/supply data and virtual enterprise visibility information in real time that will reduce uncertainty to improve decision effectiveness.

Supply chain operations that involve actions outside the enterprise's four walls require monitoring and measuring trading partner performance as never before due to virtual enterprise integration of the core competencies across the value chain. To realize this new capability, enterprises must reengineer their decision-making processes in such a manner so that they are able to adapt to the dynamics needed to synchronize supply with demand. Decision makers must be able to evaluate performance across the virtual enterprise using common metrics in order to make effective decisions for guiding value chain strategies, directing logistics tactics, and managing operational activities in a real-time manner. As c-commerce evolves as the mainstream e-business strategy for effective value chain management, the reengineering of management decision making processes becomes the critical success factor for enterprise profitability and growth in the 21st century.

C-commerce in Perspective

In 1995 something wonderful happened, something quite remarkable—a true phenomenon. During that year the whole world unilaterally adopted the same set of IT infrastructure standards to share business data, the Internet. No governmental edict or channel master command forced this technological paradigm shift. In 1994, the Internet was still an arcane technology primarily known and used by academia, computer scientists, and the military. In 1996, businesses were being established and were generating revenue from commercial buying and selling transactions by leveraging the enabling technology known as the World Wide Web. This seminal event has forever changed the way we do business—now, e-business.

There is a consensus that e-business entered the collective consciousness of the business mainstream in 1995, when the scope of traditional enterprise resource planning (ERP) business applications expanded to embrace trading partners for supply chain management (SCM) and customer relationship management (CRM). This provided the basis for Internet-enabled procurement and triggered deployment of electronic commerce (e-commerce) applications to exchange buy/sell transactions to purchase direct and indirect materials via business-to-business (B2B) supply chain scenarios, as well as finished products and customer services via business-to-consumer (B2C) channels.

C-commerce represents the evolution of e-business beyond e-commerce. Rather than simply exchanging procurement transactions, enterprises are sharing intellectual capital with their trading partners working as a value chain that provides a competitive advantage for the development and distribution of their products. These collaborative business practices are fundamental to the often stated e-business transformation. The convergence of business process reengineering and Internet technology that spawned e-business during the 1990s has set the stage for reengineering the resulting management decision making processes to promote c-commerce as the dominant business model of the first decade of the 21st century.

Note that the collaborative business practices that are fundamental to c-commerce were mature and employed as standard practices in aerospace well before 1995. In fact, the c-commerce "best practices," leveraging Internet-based collaboration in the aerospace industry, had their beginnings in 1985. This should be no surprise when it is realized that it was natural for the aerospace enterprises to collaborate over a communications network to share data with their customers in the Department of Defense (DoD) since they created the Internet. By 1994, leading aerospace prime contractors were providing business and technical applications hosted on integration hub platforms for collaborative product development and logistic support for their DoD customers, as well as for their value-added trading partners and suppliers.

This same integration hub, which is the hallmark of c-commerce architecture, facilitates the sharing of information between trading partners in a value chain that operates as a virtual enterprise. The use of integration hubs, as platforms for private trading exchanges (PTX), by Cisco, Dell, General Electric, and Intel has been widely reported over the past several years. The results of these PTX deployments have been a key factor contributing to these market leaders' ability to realize a sustainable competitive advantage. As with the technology adoption model, which consistently predicts the IT behavior of individual companies, there are also analogous Type A, B, and C industries in terms of the levels of risk tolerance for IT adoption which enables c-commerce business practices. The behavior as exhibited by such industries as aerospace and high tech electronics clearly indicates that they are Type A as pioneers of advanced technologies and aggressively adopt high-risk strategies to gain the high-potential rewards. Use of the integration hub architecture to blend intranet, extranet, and Internet activities into a seamless shared data environment is one of the trademarks of a c-commerce leader.

Such collaborative business practices have begun to migrate from intranets to extranets across a diverse base of enterprises that are market share leaders in all major industrial sectors of our global economy. A recent study, published by Deloitte Research (2002), of 356 enterprises reports that adoption of c-commerce is growing across most industries, including the public sector. This momentum is being accelerated by today's economic climate in order to improve both top-line and bottom-line business performance, where approximately 74% of the enterprises stated that c-commerce is critically important to their business. There is evidence from a recent survey of 300 business-technology executives, who are making IT investments to gain both cost savings and access to information to be able to make better decisions, that the pace of c-commerce adoption is accelerating at an increasing rate (Swanson, 2002). Collaborative initiatives are underway in 69% of the enterprises and another 16% are planning them. The survey reports that 76% of the enterprises are targeting projects with customers, 72% with trading partners, and 61% with suppliers. Such agility is also mission critical in the federal government where the DoD is accelerating its collaborative defense department initiative that will embrace the commercial "best practices" for SCM, as well as collaborative product development (Moad, 2002). It is these c-commerce business processes that form the compelling value propositions for the PTX deployments and successful e-marketplace initiatives that are building the foundation of the 21st century economy.

This convergence of enabling information technology with global marketplace economics has produced an international business climate that drives the demand by trading partners for the ability to securely create, manage, and use data between shared enterprise business processes. The concepts of the integration hub that are embodied in enterprise information portals (EIP), serving as e-business platforms that are available as packaged software products to enable c-commerce, can be utilized to achieve virtual enterprise operations now. The c-commerce strategy for integration hub architecture embraces an open systems environment, and the adoption of international standards, as well as coordination of emerging industry standards and best practices for a secure, shared data environment. The direct benefits are realized in terms of substantial reductions in product time to market and life cycle costs, as well as significant improvements in the quality and performance of the products.

Virtual enterprise management is an information intensive process that involves a value chain to design, build, and support a product throughout its life cycle. The core competencies that enable the full spectrum of c-commerce, as shown in Figure 1, are changing so rapidly that enterprises are finding it difficult to align their business plans with this new market force. This includes integrating internal business processes via intranets, optimizing external business relationships via extranets, as well as leveraging B2B models via a PTX and e-marketplaces as well as B2C channels via Internet Web sites. This wide range of information is created, managed, and shared by a large number of organizations and their trading partners in a value chain that forms a virtual enterprise. Virtual enterprise formation and operation scenarios must be based on trading partner core competencies and value-added capabilities for c-commerce.

The full spectrum c-commerce model is defined in terms of virtual enterprise integration for value chain management of trading partner core competencies. For any enterprise, the operational shared data environment consisting of the evolutionary level of maturity for c-commerce capabilities as a value-added trading partner may be defined as follows:

- Electronic data interchange,
- Enterprise integration,
- Product data interchange,
- Collaborative business practices,
- Outsourcing via e-marketplaces,
- Product development/distribution via value chains, and
- Virtual enterprise operations.

Vertical integration advantages have been overcome by the need for enterprise agility and focus on core competencies. No enterprise is sufficiently large, or financially independent, or organizationally smart enough to be world class in all capabilities needed to design, build, and support complex products throughout their life cycles. Business models formed by vertical integration strategies

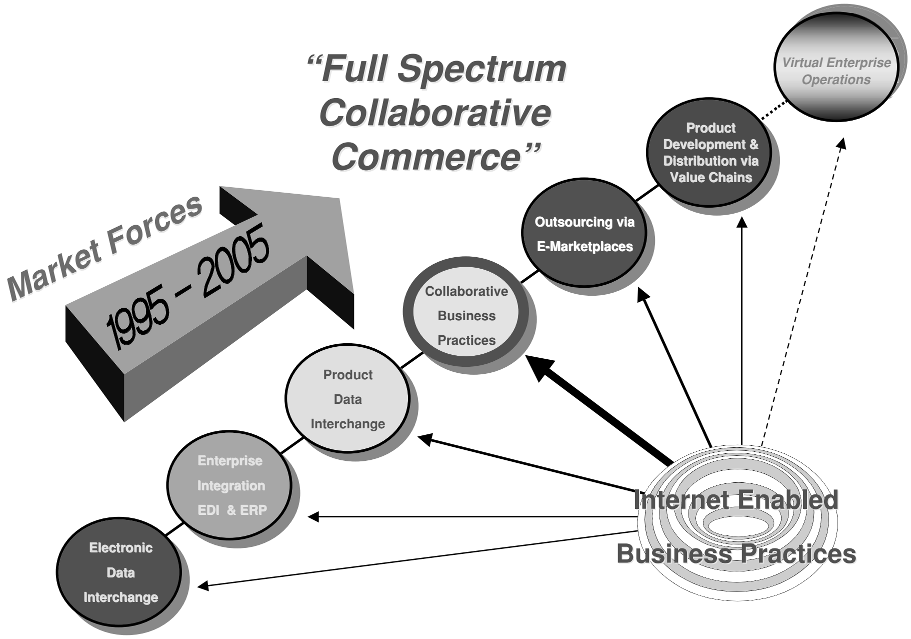

"Full Spectrum Collaborative Commerce"
Market Forces
1995 – 2005
Virtual Enterprise Operations
Product Development & Distribution via Value Chains
Outsourcing via E-Marketplaces
Collaborative Business Practices
Product Data Interchange
Enterprise Integration EDI & ERP
Electronic Data Interchange
Internet Enabled Business Practices

Figure 1: Core competencies for c-commerce.

are being transformed by convergence of information technology and global marketplace economics into e-business models based on virtual enterprise integration strategies using c-commerce practices.

Critical Success Factors

In our 21st century economy, product development agility and distribution speed are fundamentals, as are operational excellence, customer loyalty, and continuous innovation, as well as sustainable value creation. Yet most enterprises use a business structure and management decision making model inherited from the industrial age of the 19th century with barriers to these critical success factors. Furthermore while e-business strategies for operational excellence have been adopted with success during the 1990s, it may be necessary but not sufficient to achieve a sustainable competitive advantage in an age where information is replacing inventory as the key enterprise asset. Internet technology has enabled SCM tools to provide accurate and timely information on the status of inventory within the enterprise and across the value chain (Murphy, 2002).

Collaborative business practices require the sharing of business and technical data that is often proprietary or at least "competition sensitive" information. Trust and data/process integration standards are imperatives for effective value chain management. This decision making necessitates establishing business rules that are driven by value chain needs in formation and operation of the virtual enterprise based on the trading partners' risk/reward contributions and value-added core competencies. C-commerce critical success factors are

- leveraging Internet technologies for first internal and then external data sharing,
- providing loosely coupled application interoperability via integration hubs across the value chain,
- focusing on core competencies associated with the established collaborative business practices for the demand chain via collaborative product commerce (CPC) models and the supply chain via collaborative planning, forecasting, and replenishment (CPFR),
- building virtual enterprises on trusted value chains to redefine competitive advantage,
- generating real-time visibility, event notification, and performance measurement throughout the value chain, and
- reengineering value chain management decision making processes.

C-commerce strategies are being built around CPC and CPFR business models. While CPC scenarios are most prevalent in the industrial products sector for build-to-order solutions, CPFR was created to solve supply chain problems in the consumer packaged goods (CPG) market space. Leading enterprises are employing both CPC on the demand side and CPFR on the supply side to form the basis of the value chain management decision making.

CPFR was initiated and trademarked in 1998 by the Voluntary Inter-industry Commerce Standards (VICS) association, a nonprofit organization taking a global leadership role in streamlining the flow of product information throughout the retail value chain. This innovative business practice is becoming a leading B2B model for c-commerce in the CPG industry. It uses forecasting models and real-time point-of-sale transaction information to produce a collaborative environment for trading partners to improve efficiencies in inventories, customer service, and bottom-line savings. CPFR facilitates collaborative relationships between buyers and sellers through co-managed processes and shared information. The goal of CPFR is to drive increased interoperability between trading partners so that their different information systems are able to seamlessly share information to realize more effective decision making. The benefits for all of the trading partners in using CPFR are to improve process efficiencies, increase sales, decrease fixed assets and working capital, and reduce inventory for the entire value chain while satisfying consumer needs by synchronizing demand and supply information (VICS, 2002).

Some of the biggest names in retailing such as Proctor & Gamble (P&G), Sears, and Wal-Mart have been early adopters of CPFR, with P&G, as well as Unilever Group and Kimberly-Clark, managing up to half their product lines via CPFR. Results to date are impressive: inventory including safety stock reductions of 10 to 20%, point-of-sale stock out reductions of 2 to 4%, with inventory turnover rates doubled. Heineken has the longest running CPFR initiative with a SCM application that supports 450 distributors by reducing order cycle times from three months to four weeks (Bowman, 2002).

These market leaders have been realizing significant benefits by deploying CPFR business practices via a PTX. CPFR has matured and proven its value proposition as an enabler of c-commerce with analyst estimates of $40 billion a year savings throughout the CPG industry resulting from driving out inaccuracies in information, inefficiencies in processes, as well as reductions in excess inventory that serves as safety stock at each stage of the value chain. Now several e-marketplaces that facilitate CPG value chains have embraced the UCCnet Global Registry data standards as a common language for collaborative SCM dialog (Swanson, 2002; Konicki, 2002). Worldwide Retail Exchange, Transora, and GlobalNetXchange have recognized that the inability to share information in an unambiguous manner has been a barrier to success. P&G is working with Transora to host and incrementally implement its 60,000-item product catalog as it executes a c-commerce strategy as an enterprise CPFR scenario. UCCnet provides links from the registry data to the P&G catalog on Transora to facilitate collaborative ordering, as well as transportation and inventory management activities with retailers.

Although less formally organized than CPFR, demand chain integration is already being realized in many Type A industries as popularized by Dell's B2C success story in the computer industry, using a CPC business model with scenarios such as B2B collaborative product development, build-to-order for mass customization, and collaborative sourcing (Halpern, 2001). CPC is the c-commerce strategy that spans the product life cycle with value chain activities that have reengineered the product definition, development, and delivery business processes. It involves promoting product data as intellectual capital for a shared asset of a virtual enterprise. The adoption of CPC principles is making new product introductions better, faster, and cheaper from such diverse groups as the over 80 companies of the Lockheed Martin Joint Strike Fighter team working in a virtual enterprise to build the new family of stealth aircraft for the United States and its allies to Microsoft and its manufacturing partner Flextronics to bring the new Xbox video game product to market (Keenan & Ante, 2002).

The integration hub architecture is the fundamental e-business platform to enable both CPC and CPFR value chain scenarios. Virtual enterprise data management services are contractually required for trading partner information as part of a service level agreement. They represent the c-commerce value proposition for integration and interoperability success to be realized whether provided by a PTX or e-marketplace. This services-oriented architecture, depicted in Figure 2, supports a virtual enterprise data management model via Web services that facilitate a standards-based, secure, shared data environment enabled by business process workflow technology for value chain management.

The integration services provided by EIPs act as virtual enterprise gateways providing semantic interoperability with diverse applications across multiple PTXs and public e-marketplaces. An integration hub contains a profile of every registered virtual enterprise member, including a catalog of their products, capabilities, and capacities, as well as their trading partner agreement and communication system preferences. This profile provides data directory and dictionary services that describe all the sharable information at the data element level that is owned by a

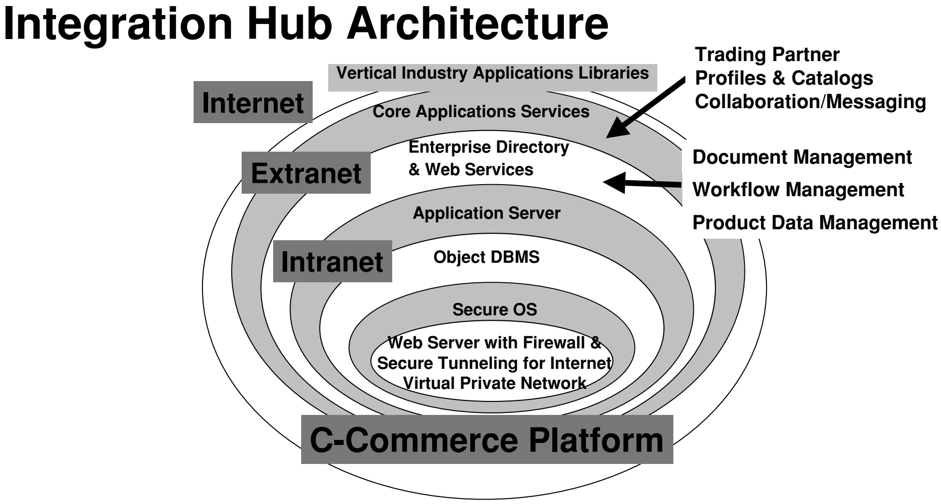


Figure 2: Integration hub data management services.

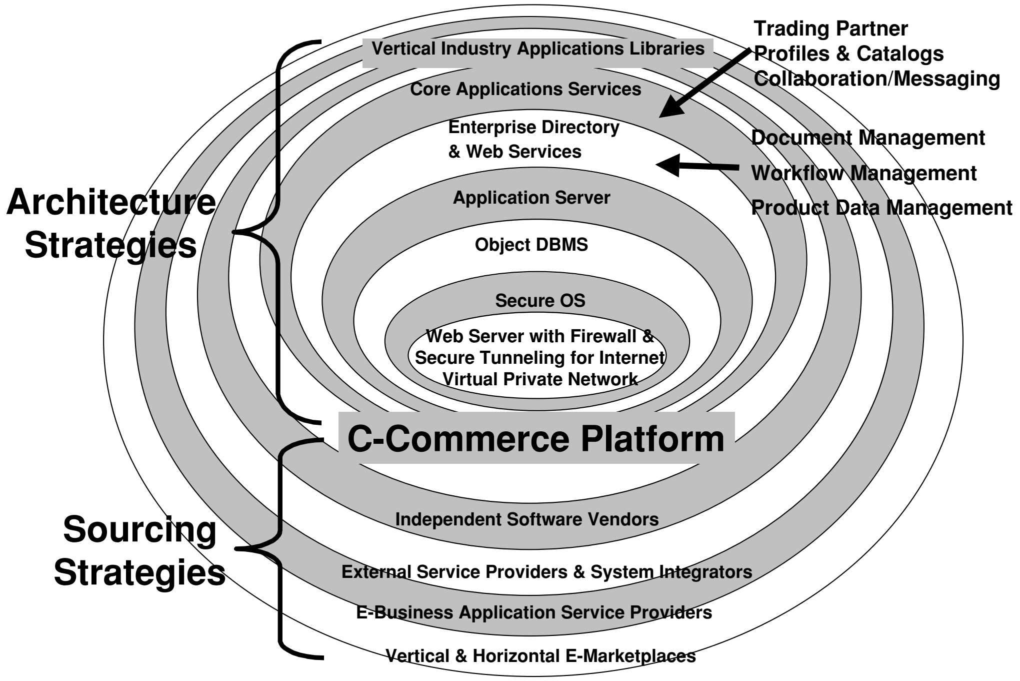

Vertical Industry Applications Libraries
Core Applications Services
Enterprise Directory & Web Services
Application Server
Object DBMS
Secure OS
Web Server with Firewall & Secure Tunneling for Internet Virtual Private Network
Trading Partner Profiles & Catalogs Collaboration/Messaging
Document Management
Workflow Management
Product Data Management
Architecture Strategies
C-Commerce Platform
Independent Software Vendors
External Service Providers & System Integrators
E-Business Application Service Providers
Vertical & Horizontal E-Marketplaces
Sourcing Strategies

Figure 3: Integration hub architecture and sourcing strategies.

trading partner, including a security policy to permit the sharing of information with a virtual enterprise customer, but not with a potential competitor.

Virtual enterprise data management services include data format and communication translations as required for messaging services. Collaboration tools can also be provided to support both asynchronous groupware and synchronous real-time, multipoint data conferencing applications for both structured and unstructured data. In addition, an integration hub can offer product data exchange translation in accordance with the governing industry standards, as well as a library of application software for a vertical industry where native data format exchange is considered to be mission critical, such as the aerospace industry.

Core platform services, based on a services-oriented architecture embracing open systems concepts, are hosted on a Web server with firewalls and secure tunneling capable of handling public key encryption transactions. Layered Web services extend outward to support directory and middleware, including interoperable object management services for workflow, documents, and product data; generic business collaboration tools and messaging services; trading partner profile and catalog services to facilitate virtual enterprise partnering; and customizable vertical industry specific tools and application services.

The messaging services provide the enabling technology that is fundamental for decision making in a real-time enterprise. Such c-commerce messaging data provides the status of stock-on-hand that is stored at enterprise and trading partner locations, as well as in-transit across the value chain for inventory visibility. Furthermore, business activity monitoring (BAM) solutions generate event notifications associated with the business rules that reflect the roles and responsibilities of each trading partner. Finally, operational metrics are displayed as real-time dashboards for measurement of trading partner performance in accordance with their service level agreements to support value chain management decision making.

An EIP deployment, using the integration hub sourcing strategies shown in Figure 3, may be outsourced to a trading partner with an IT core competency in order to allow the enterprise to focus on executing its strategic business plan. These integration services will offer virtual enterprise gateways providing semantic interoperability with diverse applications such as a PTX that must operate with several intranets and extranets across multiple vertical industries as well as several public e-marketplaces. In general, enterprises need to build or acquire access to networks outside their four walls that will interoperate as a blend of extranets that consist of PTXs, as well as both public vertical industry and horizontal e-marketplaces that provide business process outsourcing and information integration services such as Exostar serving the aerospace industry and E2open providing supplier onboarding for high-tech electronics value chains.

The imperatives for these extranets can be best understood by emphasizing the need for c-commerce business rules. Enterprises must base these rules on the virtual enterprise data management model in a manner that reflects the unique collaborative needs for the specific set of trading partner networks that support the operation of the virtual enterprise.

In summary, c-commerce critical success factors include leveraging Internet technologies for first internal and then external data sharing; providing loosely coupled semantic interoperability for enterprise applications via integration hubs across the value chain; focusing on core competencies for collaborative business practices;

and building virtual enterprises on trusted value chains to redefine competitive advantage, generating real-time visibility, event notification, and performance measurement throughout the value chain, as well as reengineering value chain management decision making processes.

Future Trends

The e-business transformation that is being realized via c-commerce can be described in terms of management decision making processes utilizing feedback to leverage real-time virtual enterprise information. The two key adaptive strategic planning processes are

Building the virtual enterprise infrastructure using integration hubs, and

Making decisions to optimize value chain business performance.

As previously stated integration hubs, deployed as platforms for either a PTX or e-marketplace, provide the environment that facilitates c-commerce Web services for applications such as inventory visibility, BAM, and performance metrics displayed in real-time dashboards. Identifying and analyzing value chain performance trends enables enterprises to be able to adapt their strategic plans, redesign virtual enterprise configurations, and reallocate resources for their core competencies in order to optimize profitability in response to customer and market dynamics (Hoffman, 2002). The benefits of c-commerce can then be realized via leveraging the Internet as an enabling technology in order to transform enterprise core business management decision making practices.

This trend has renewed business interest in the *Balanced Scorecard* as part of a strategic business planning methodology that is being used to communicate the enterprise business strategy throughout the value chain. Kaplan and Norton (2001), the creators of the balanced scorecard, describe adaptive strategic planning in the context of "making strategy a continual process." Accordingly, there are three key e-business drivers for including the Balanced Scorecard as a fundamental model for the solution to the c-commerce strategy management problem:

Linking strategy and budgeting: Strategic initiatives and performance targets on the Balanced Scorecard link the enterprise strategy to the specific resource allocations for operating each value chain using "rolling forecasts" in a manner that reflects operating in continuously changing environments.

Closing the strategy loop: C-commerce messaging applications including systems displaying performance metrics in real-time dashboards can be linked to the Balanced Scorecard providing a new framework for exception reporting and adaptive business management practices, as well as focus on intangibles with leading indicators, such as intellectual and social capital, management talent, core competencies, and customer loyalty assets that are neglected by traditional lagging financial measurement.

Testing, learning, and adapting: Virtual enterprise management can be more proactive by testing and evaluating strategic value chain hypotheses with the information from the c-commerce messaging applications. Strategies for each value chain can evolve in real time in response to market dynamics.

An adaptive strategic planning system enables value chain management teams to make strategy a continual process by monitoring performance against enterprise and line-of-business strategies, updating the performance measures on the balanced scorecards using actual operations data in real time, collaborating as virtual teams with an executive dashboard to interpret the performance metrics, developing strategic insights with knowledge management tools in order to formulate new competitive directions, and then redetermining enterprise resource allocations to continuously reflect the dynamics of the business environment.

Such a c-commerce planning and control application, implemented as a decision support system (DSS), can be based on an enterprise and line-of-business Balanced Scorecard as well as real-time SCM performance analytics. It enables progress in executing the strategy to be monitored, as well as supports corrective actions to be performed in a timely manner. This adaptive strategic planning system serves as an enabler of a c-commerce strategy improvement process that links the real-time value chain performance, as feedback for the business operations control process, with the enterprise-level strategy.

The most complete and widely disseminated framework for the reengineering of the management decision making process is the *Beyond Budgeting Round Table* (BBRT) model that was developed by the Consortium for Advanced Manufacturing International (CAM-I), a not-for-profit collaborative research group. Sixty enterprises have sponsored the work of the BBRT since it was launched in 1998 to investigate organizations that had replaced their traditional centralized "budgeting" management model with an alternative model. The BBRT model is based on decentralized business decision making and adaptive management practices. The BBRT research shows that fixed strategies prevent agile responses, rigid organization structures inhibit innovation, entrenched functional silos undermine collaborative processes, focus on product targets works against customer loyalty programs, and short-term performance contracts fail to support long-term value creation. Implementing the BBRT model involves introducing new performance management processes, as well as decentralizing decision making power. As a result of embracing BBRT, BP-Amoco and Asea Brown Boveri have integrated budgeting as part of an adaptive strategic planning process, while Ford, Electrolux, and Sprint have leveraged their information infrastructure to enable real-time performance measurement for a "rolling budget" process (Bunce, Fraser, & Hope, 2002).

CAM-I believes that the BBRT initiative has a generic applicability. BBRT research confirms that operational excellence is no longer a sufficient element of a business strategy, although as with total quality, it is still an essential element of enterprise operating capabilities. Therefore, only enterprises that pursue continuous innovation or customer intimacy strategies that employ decentralized and adaptive decision making processes will prosper in the 21st century.

STRATEGIC PLANNING FOR BUILDING C-COMMERCE

Throughout the global marketplace, executive teams are asking the same questions that show concern over enterprise sustainable competitive advantage and even survival:

How are my industry's business models changing?
What candidate business models will work for me?
Where are the gaps in my enterprise readiness?
What road map can I use to manage these changes?

The challenge is to exploit the opportunities presented by c-commerce and then determine what and where the level of investment should be in the enterprise IT portfolio.

C-commerce Strategic Planning Methodology

Enterprises may employ a c-commerce strategic planning methodology to build a road map for the design, development, and deployment of a virtual enterprise enabled by real-time IT services-oriented architecture, as follows.

Step 1: Visioning via the Strategy Framework

During this step, the enterprise must establish a corporate e-business vision, and therefore, a resulting high-level c-commerce strategic direction as to how the enterprise is "To-Be" in the envisioned future. A top-down and bottom-up approach is adopted in Step 1 to ensure that the entire organization has input into establishing the corporate e-business vision of the "To-Be" enterprise and buys into it with CEO leadership.

Strategic CPC and CPFR scenarios using an integration hub services-oriented architecture reflect the different applications of value chain management under four main categories:

Demand chain management,
Customer/partner relationship management,
Supply chain management, and
Supplier relationship management.

Any c-commerce initiative can be positioned according to one of these four headings. Having identified the four ways in which c-commerce can be applied, strategic scenarios are ranked through a robust evaluation process in terms of the value of opportunity each scenario represents to the enterprise as well as the ability to deploy it. The resultant ideal "target zone" can be represented diagrammatically as illustrated in Figure 4.

Enterprises must formally evaluate each of the candidate scenarios against a number of business factors. These business factors relate either to "opportunity" or "capability" and are scored under each heading which generally consists of from 8 to 12 metrics that reflect the enterprise balanced scorecard.

The c-commerce visioning process requires three perspectives. The enterprise perspective provides insight of a company's current capabilities, culture, and vision. The marketplace perspective provides a view of realities and changes in industry practice, barriers to entry, competition, constituent behavior, and offerings. The technology perspective, with its associated policy impacts on the enterprise and marketplace, extends strategic thinking "out of the box" by discounting commonly held beliefs about markets, products, and competition. This forces enterprises to address new methods of doing

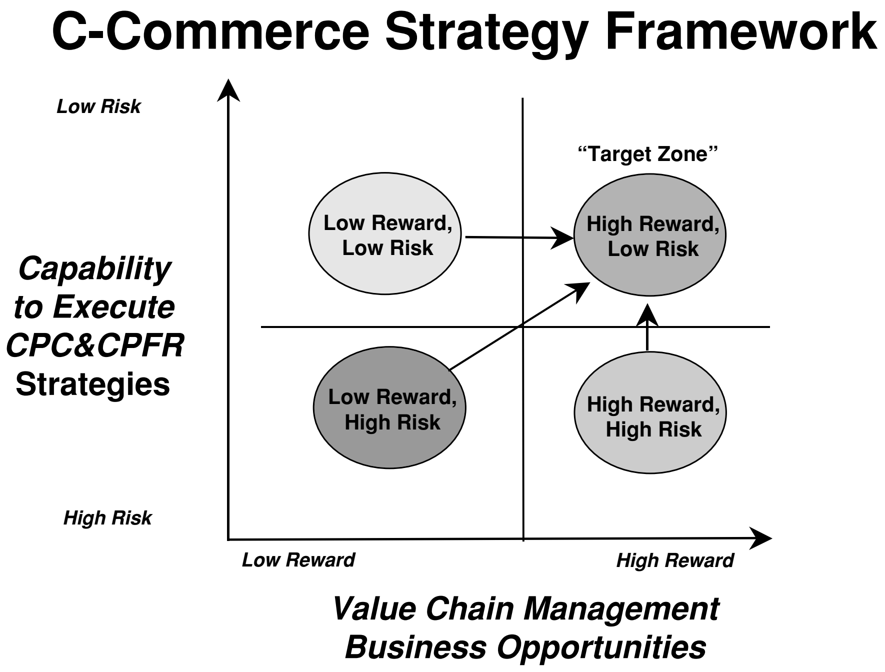


Figure 4: Strategic visioning for c-commerce.

business, new product offerings, and new competitive paradigms.

Development of the e-business vision and c-commerce strategy is not a one-time process. Because of the evolving nature of best practices, state-of-the-art product offerings, quality expectations and cultural changes, strategy development is iterative. C-commerce business models and strategies must be updated on a continuous basis.

Step 2: Virtual Enterprise Readiness Assessment

In Step 2, existing or "As-Is" enterprise processes and infrastructure are formally evaluated in order to establish a baseline of enterprise capabilities. The existing processes will indicate the need for and the ability of the organization to adapt to change. The infrastructure assessment will provide an understanding of the level of investment needed to implement a c-commerce initiative. The combination of these metrics will provide a score that will allow the aforementioned CPC and CPFR scenarios to be positioned in the c-commerce strategy framework.

Enterprises need to assess the maturity of their business, technology, and organization domains for evaluating c-commerce readiness (Sabath & Fontanella, 2002). They must assimilate the information gained through a due-diligence process for assessment of their core competencies and develop a mapping of pertinent IT capabilities to c-commerce business practices. This process should include a review of enabling technology and business model trends in their industry as a benchmark. Solidifying the understanding of the enterprise's technology and business challenges as well as developing the knowledge of core competencies concerning technology, business processes, and organizational capabilities is fundamental to crafting a c-commerce business strategy. The enterprise can then build a trading partner value proposition that can be leveraged in forming or joining virtual enterprises in a way that make strategic business sense.

Enterprises should look at three technology categories to assess their readiness to support a c-commerce business model: application portfolio, systems infrastructure, and network infrastructure. They should compare their "collaborative business vision" to benchmarks pertaining to B2B and B2C e-business models, as well as industry trends for CPC and CPFR value chain scenarios. It is critical for enterprises to evaluate their business process infrastructure, as well as organizational structure with respect to their ability to support customer-centric value chains. Again it is essential for enterprises to understand their core competencies from a product life cycle perspective, in terms of product development through product support, including customer and supplier relationship management capabilities.

Step 3: C-commerce Gap Analysis

Although the preliminary strategies and initiatives were identified in light of both enterprise opportunities and capabilities there is still a need to subject the enterprise c-commerce strategic architecture to a formal "gap analysis." In this step of the c-commerce strategic planning methodology the magnitude of the difference between the current "As-Is" environment as assessed in Step 2 and the future "To-Be" state of the enterprise as envisioned in Step 1 is formally evaluated.

A complete gap analysis, at a corporate level and in relation to each line of business must cover the following four elements:

- E-business drivers,
- Business processes,
- Enabling technologies, and
- Organization and culture.

This gap analysis must precede the next step of the project, which is to make recommendations on the implementation of the strategy. Performing the gap analysis allows for the identification of specific management approaches or activities required ensuring that the lines of business can execute their strategies as planned in support of the corporate business model.

Step 4: Strategic VEI Architecture

Virtual enterprise integration (VEI) is conducted in Step 4. This is where the vision, strategies, and initiatives generated during Step 1 are synthesized into an enterprise c-commerce strategic architecture. This top-level enterprise scenario can then be executed by the operational lines of business and integrated into their c-commerce strategy process with their resulting line of business level Balanced Scorecard performance management system.

The enterprise c-commerce strategic architecture integrates the total of all the initiatives developed within the corporate-level visioning and business-line-level initiatives and collectively represents the associated corporate and lines of business strategic plans. The strategic architecture for VEI comprises

- The e-business vision for c-commerce strategies;
- The core strategies applicable at a corporate level and a business line level, and their prioritization based on the attractiveness blend of the opportunity values and capabilities risks; and
- The c-commerce initiatives that will realize the strategies at corporate- and business-line levels, as well as their relative prioritization.

Note that this concept of VEI utilizes a three architecture approach as a model for c-commerce initiatives that provide real-time decision making for value chain management:

- VEI technology architecture for integration hub platforms via EIP products and Web services provides semantic interoperability for value chain visibility;
- C-commerce application architecture for value chain management integration of demand chain management, customer/partner relationship management, supplier relationship management, and supply chain management provides real-time BAM for event notification; and
- E-business process architecture for CPC and CPFR scenarios provides adaptive performance measurement.

Step 5: Implementation Road Map via Value Chain Analysis

The results of the gap analysis and strategic architecture development must be sequenced and the interproject dependencies must be taken into consideration in order to develop a road map for how to resolve the identified gaps. The implementation recommendations will also provide inputs into the strategic planning process that exists at a line of business level because each of the business lines may need to adopt enterprise recommendations for implementation of their specific initiatives and projects.

It is important for business leaders to understand that pilot testing of emerging collaborative business practices, in a representative management decision-making environment that has been instrumented for business case analysis, is the only rational way to mitigate the change management risks associated with the execution of a comprehensive enterprise strategy. An implementation road map must be orchestrated to identify and evaluate potential business process and enabling technology gaps by use of such a shared data environment capable of supporting the intranet, extranet, and Internet value chain scenarios, as well as to deliver incremental benefits in a self-funding or pay-as-you-go manner. The pilot test and evaluation results will also provide implementation requirements for recommended information technology investments. The pilot testing will further enhance the enterprise c-commerce strategic planning process that exists at a business line level (e.g., value chains for key commodities or customer/trading partner initiatives).

Incremental Implementation Road Map for C-commerce Pilot Projects

Execution of a comprehensive enterprise c-commerce strategy necessitates an endeavor that consists of enabling value chain initiatives and their related implementation road map. The following series of tasks serve as a guideline for navigating the incremental road map to implement the c-commerce pilot projects:

Task 1—Develop c-commerce business model by value chain.

Task 2—Develop e-business services architecture.

Task 3—Define virtual enterprise formation and CPC/CPFR scenarios.

Task 4—Develop intranet/extranet systems configuration.

Task 5—Perform intranet/extranet systems integration.

Task 6—Test c-commerce environment.

Task 7—Deliver lessons learned and business case analyses.

This pilot project incremental implementation road map is a framework for creating and deploying value chain management scenarios. It is further defined in terms of representative EIP service-oriented architecture products which may be used to source a scalable and extensible integration hub platform. Key to an effective EIP product selection is a solid understanding of enterprise requirements and identification of the audiences for which the portal will be targeted. Different audiences (employees vs. suppliers vs. trading partners vs. customers) will have very different needs, tastes, demographics, and requirements. Multiple portals in any large enterprise are quite likely and even desirable, since the EIP product selected must meet the needs of each intended community of interest for their intranet, extranet, or Internet-based value chain management scenarios. Choosing a product requires sound research and expert advice. Enterprises should leverage all available technology resources, perform due diligence on comparable reference implementations, and conduct pilot tests with the EIP product vendors before the selection of the best source for each e-business platform is finalized.

Using the incremental implementation model as a framework for a pilot project road map, enterprises can develop c-commerce competencies by investing in a portfolio of IT initiatives, which may include the following examples:

1. Integration hub infrastructure development via e-business platform,
2. Intranet integration via integration hub,
3. PTX for CPC extranet integration via integration hub, and
4. E-marketplace membership for CPFR Internet integration.

The value of the first project is to acquire the skills and know how to establish a baseline c-commerce capability and reusable enterprise integration technology platform that will meet the enterprise's present and future e-business needs.

The second project will provide a single, unified face to employees, suppliers, and customers as trading partners, worldwide. These intranets enable the enterprise's business lines to organize value chains in response to market dynamics. This results in the realization of virtual enterprise benefits, such as the integration of multiple ERP systems for seamless product transfer from distribution centers located across global operations in order to accommodate localized customer requirements and regional demand spikes. Such a virtual enterprise environment facilitates transformation of low-margin commodities into higher margin products that are differentiated by value-added services and mass customization value propositions.

An extranet implementation, which necessitates semantic interoperability for a more advanced degree of c-commerce maturity, serves as the third project using CPC business models. They employ an integration hub with product development and distribution functionality that facilitates each value chain as appropriate. This will enable the enterprise to control the specification and design of products for mass customization. CPC is a c-commerce strategy for exploiting new Web-enabled opportunities across product life cycle processes. Opportunities include both B2B and B2C business models, such as collaborative product development, customer self-service design, strategic product sourcing, manufacturing/distribution collaboration, and product support self-service portals.

The fourth project focuses on outsourcing via an e-marketplace membership to acquire the trading partner integration services needed by CPFR business models. Such a scenario leverages the public Internet for connectivity and Web services with virtual private networking capabilities to provide a secure shared data environment. These virtual enterprise data management services may be employed on a fee for service basis as appropriate for sharing contractually required trading partner information.

In summary, strategic planning for building c-commerce can incorporate a Balanced Scorecard utilizing a real-time performance management system to facilitate an adaptive strategic planning process for value chain optimization. An individual business case for each c-commerce initiative must be orchestrated in a manner that is logically consistent with the respective enterprise and lines of business balanced scorecards. This c-commerce strategic planning methodology blends CPC and CPFR capabilities in value chain scenarios to provide the basis for a compelling customer value proposition, as well as to realize VEI benefits for a sustainable competitive advantage. The integrated IT investment portfolio of c-commerce initiatives consists of strategic value chain pilot projects. The pilot projects may build on recommendations that enterprises employ a technology approach that evolves c-commerce capabilities to improve overall enterprise readiness and integrate key technical and application architecture components in a customer-centric manner via CPC and CPFR scenarios. The projects may produce pilot value chain solutions that can realize immediate competitive advantage and bottom-line payoff. The portfolio of CPC/CPFR pilots may then serve as a critical e-business transformation model for virtual enterprise change management, as well as provide the fundamental incremental investments for self-funding of c-commerce initiatives.

STRATEGIC PLANNING FOR PERFORMING C-COMMERCE

A challenge in the new 21st century economy is to adapt quickly to new business opportunities or threats by realigning enterprise strategy to efficiently and effectively create value. An adaptive strategic planning solution for enterprise business management decision making can provide a holistic approach to strategy management, business planning, target setting, rolling and event-driven forecasting, and business performance management based on financial and leading nonfinancial key performance indicators (KPIs). DSS applications can be used to support these activities in executing c-commerce strategies across a virtual enterprise that consists of multiple value chains with diverse organizational structures and management decision making processes.

Adaptive Strategic Planning Framework

Enterprises need a framework that facilitates adaptive strategic planning for value chain management. A suite of DSS applications provides the functionality needed to implement c-commerce business models. A continuous c-commerce strategic planning process approach, which adapts to the real-time execution of the business plans, is essential to value chain management success.

The four key DSS framework elements deployed in a service-oriented architecture as a network of integration hub platforms are as follows:

Execute via Balanced Scorecard,
Measure via Business Activity Monitoring,
Analyze via Knowledge Management, and
Model via Dynamic Simulation.

Execute via Balanced Scorecard

The execute function allows an enterprise to continuously translate c-commerce strategy into action within day-to-day business activities and to constantly adapt strategy with real-time feedback based on the actual operational performance of each value chain. Adaptive ability can be expressed in terms of three types of actions resulting from the management decision-making process: predictive, proactive, and reactive. Predictive controls involve anticipating a problem and deploying a solution in time to avoid the adversity. Proactive responses are associated with smooth recovery from a problem before it escalates. Reactive capabilities are limited by supply chain uncertainties to just-in-case deployments of inventory safety stock, as well as excessive resource allocation for production and distribution capacity in order to maintain contractual service level agreements.

Measure via Business Activity Monitoring

The most important feature of the adaptive strategic planning process is the ability to provide feedback by means of real-time performance measurement. As the enterprise executes its c-commerce strategy for each value chain, the world around it changes and that strategy must adapt accordingly. As a strategic management practice, adaptive strategic planning must be a learning process. The DSS should facilitate a dialog throughout the value chain so that performance can be interpreted, knowledge can be transferred, and improvements can be made. Real-time measurement is crucial because the longer the BAM latency the more management is limited to reactive rather than proactive decision making. Leading enterprises are using IT solutions for SCM to speed relevant performance information to the appropriate managers and to display these KPIs in a desktop dashboard as exception reporting to focus decision making attention. Critical success factors involve not only the determination of the correct KPIs, but also the refresh frequency for each KPI on each manager's dashboard.

Analyze via Knowledge Management

Key enabling technology trends in knowledge discovery and business intelligence are making management decision making breakthroughs possible. Data mining techniques permit extracting and transforming of data from multiple sources. Data mining can be classified as either descriptive or predictive, where descriptive mining tasks characterize the general properties of the data while predictive mining tasks perform inferences on the source data. These knowledge discovery techniques can be used to explore trends and relationships across a value chain. Knowledge management enabled by a DSS with these tools and databases can support analytic approaches used to test and evaluate value chain design and operational policy hypotheses, explain the higher level trends associated with the dynamics of supply and demand as part of the value chain management system, or generate new insights that update and refine the c-commerce strategy.

Model via Dynamic Simulation

The use of dynamic simulation of value chain business models will dramatically increase the effective execution of c-commerce strategies. Simulation allows the Balanced Scorecard cause-and-effect linkages of the strategy to be described mathematically and used for testing scenarios. This capability will help companies evaluate "what if" scenarios. It will allow the entire management team to participate in interactive sessions for the real-time development of strategy. Dynamic simulation software will have the same impact on strategic planning that spreadsheet software has had on financial planning.

The DSS for adaptive strategic planning can be applied to a range of business process modeling applications and work with live operational data to dynamically model processes. Such functionality provides a tool that allows users distributed across a virtual enterprise to collaborate in examining the potential opportunities or threats to planned business operations. The result of such collaboration is a clear understanding of the options, the risks, and the impacts across each value chain. Such predictive analyses can be shared across the value chain for joint decision making on the optimal alternatives that facilitate collaborative planning and forecasting activities.

DSS technology is emerging to enable the design and management of complex value chains for strategic, as well as tactical and eventually operational decision making. For example, at a recent Supply-Chain Council seminar (2002), Intel described how it uses a DSS at the strategic level to design a value chain based on validated CPC/CPFR models that ensure that the right product is placed in the right locations in the right amount at the right time based on service level agreement performance objectives. Then the tool can be used tactically to determine if these inventory management decisions are still valid based on the demand dynamics of the marketplace or on daily replenishment priorities, such as adaptive supply chain execution decisions in terms of which distribution center should fulfill an order.

The result is a new type of management tool that helps enterprises quickly identify the impact of changing market conditions on value chain performance and to access the best course of action to minimize risk and maximize revenue based on specified KPIs. It also provides rapid "what if" analysis to make decisions quickly and then can facilitate effective communication of that predicted performance information throughout the virtual enterprise using a service level agreement framework.

Deploying Decision Support for Adaptive Strategic Planning

An incremental implementation road map for a DSS leverages the practical application of various enabling technologies for an adaptive strategic planning framework discussed in the previous section. The fundamental concept is to generate/evaluate a strategic plan that adapts to the tactical and operational-level performance changes based upon the feedback from KPIs.

Such a road map for developing an adaptive strategic planning infrastructure is navigated in accordance with the following tasks:

1. Develop c-commerce business models by value chain. Craft a set of integrated intranet, extranet, and Internet value chain scenarios that represent supply chain optimization opportunities using the leverage of the CPC and CPFR models as a benchmark.
2. Focus on the integration hub service-oriented architecture to evaluate EIP product functionality for providing content management, enterprise application integration, and semantic interoperability services to support a DSS. Such a supply chain integration framework includes a modeling facility with value chain models that are hosted in a model repository featuring analysis and reporting tools that leverages data mining capabilities focusing on integration hub functionality for providing knowledge management capabilities.
3. Identify specific value chain operating scenarios and supply chain management issues within key enterprise initiatives and model the core processes using the dynamic simulation tools. The value chain models are populated with Balanced Scorecard KPIs. These models are then used to identify supply chain bottlenecks and analyze areas for process improvement, evaluate the business case associated with c-commerce initiatives, and design optimal value chains for specific target customer or market initiatives.
4. Focus on developing and testing the feedback mechanisms for Balanced Scorecard KPIs reporting of real-time BAM visibility data and event-based notifications captured from operational transactions that are facilitated by integration hubs. This is a key element of the adaptive decision making process that links supply chain performance management and business operations budgeting with enterprise-level c-commerce strategy.
5. Deployment of this initial adaptive strategic planning DSS completes the initial effort. The DSS facilitates changes to the alternative supply chain systems design and experimentation with potential outcomes as a result of these scenario changes, thereby "evolving" the strategic plan. The DSS provides a knowledge base for investment portfolio management of the c-commerce initiative information needed for the strategic and business planning effort, as well as facilitates the development of business cases supporting these initiatives.
6. For the adaptive strategic planning decision making process reengineering to be fully realized, the successful DSS pilot system implementation must be extended

and rehosted on an enterprise-class server. This provides the scalable integration hub platform needed to support a fully operational competency center for real-time value chain management as part of a multiphase iterative implementation effort to realize incremental benefits. The strategic Balanced Scorecard KPIs are then deployed using a distributed network of integration hubs thereby instrumenting each of the various c-commerce value chain scenarios with the appropriate dashboards. Each dashboard reflects actual KPIs as the performance metrics associated with the individual decision maker's view of the virtual enterprise operations in an actionable way.

The combination of a DSS for strategic planning with an operational value chain management system competency center provides the feedback system to optimize value chain investment, as well as assure virtual enterprise agility. Thus, the benefits of VEI are achieved by implementing an orchestrated choreography of CPC and CPFR scenarios executed in accordance with the appropriate intranet, extranet, and Internet business case analysis.

C-COMMERCE LESSONS LEARNED AND NEXT STEPS

Reengineering the business decision making process is the e-business transformation that is driven by c-commerce. This transformation is already underway by market share leaders in all major industries so that the real-time enterprise is fast becoming a reality (Siegele, 2002). It will be further accelerated by realization of three key issues:

1. Development of intellectual and social capital performance management systems that enable value chain management business case analysis methodologies to evaluate innovation and trust competencies for virtual enterprise formation, especially e-marketplace value propositions in order to facilitate small and medium size enterprises as trading partners.
2. Dissemination of process and data standards, such as UCCnet for retail products, as well as RosettaNet for electronics and ebXML for international trade, to facilitate application integration, as well as shared ontologies and semantics mediation/mapping tools for ubiquitous trading partner interoperability.
3. Creation of value chain management solutions based on Web services technologies for plug-and-play products to support dynamic VEI environments, such as generic "Collaborama" platforms capable of supporting a wide variety of diverse scenarios, including network applications from multiplayer interactive game entertainment to global environmental disaster recovery problem solving.

In conclusion, the two major incentives for trading partners to adopt c-commerce as a business strategy, productivity and competitiveness, have the value chain management model as a common denominator. Continuous productivity improvements are focused on internal enterprise integration endeavors driven by intranet applications and reengineered business practices to decrease costs and increase inherent quality. Sustainable competitive advantages are focused on external enterprise integration opportunities driven by extranet applications and newly created virtual enterprise business processes enabling collaboration to decrease time to market and increase agility. Although competitiveness can be a much greater incentive than productivity, enterprises that have not yet integrated their intranet applications and associated business processes to facilitate collaboration must do so before risking extranet deployment. The business experience and technical infrastructure that an enterprise gains from internal process collaboration and information integration initiatives provide the prerequisite capabilities that extend to trading partners in order to realize value chain management success.

GLOSSARY

Balanced scorecard A strategy and management system that provides a holistic view of enterprise success from the perspective of financial, customer value, business process, and innovation key performance measures that include both leading and lagging indicators linked in cause-and-effect relationships.

Collaborative commerce (c-commerce) A strategy for the next stage of electronic business evolution with business practices enabling trading partners to create, manage, and use data in a shared environment to design, build, and support products throughout their life cycles, working separately to leverage their core competencies together in a value chain that forms a virtual enterprise. Collaborative planning, forecasting, and replenishment (CPFR) and collaborative product commerce (CPC) are the two business collaboration patterns that have achieved broad industry consensus for realizing c-commerce strategies by enabling virtual enterprise integration. CPC applications facilitate mass customization scenarios via demand chain integration while CPFR applications facilitate mass production scenarios via supply chain integration.

Collaborative planning, forecasting, and replenishment A business practice that facilitates mutual risk/reward relationships between buyers and sellers via shared processes and information by synchronizing demand chain and supply chain activities.

Collaborative product commerce A business practice that facilitates product development between trading partners in a value chain by sharing product assets and intellectual property based on mutual business interests.

Decision support system A business management system that employs a suite of application software packages as tools, such as analytical processing and dynamic simulation, designed to assist human decision makers in solving complex problems.

Demand chain The community of organizations with the assets, processes, and information that generate elements of demand for a product.

E-marketplace An environment that operates to aggregate buyers and sellers in various vertical and horizontal industries, offering market intelligence to both parties to understand buying behavior and improve e-procurement processes, as well as collaborate with trading partners in the value chain for increased visibility to reduce inventory and processing costs.

Enterprise information portal A platform that supports targeted user communities for access to relevant structured transactional information, aggregated business intelligence data and trends, unstructured documents and collaborative sources, Internet content, and Web services, as well as interaction with integrated applications and business processes for personalized communications between enterprise trading partners.

Enterprise resource planning An integrated business management system that uses multifunction application software packages designed to serve and support multiple business functions such as planning, manufacturing, sales, marketing, accounting and finance, and human resources. The fundamental ERP platform has been extended using an open architecture with vertical industry-specific functionality to extend the enterprise into the demand chain for customer relationship management and into the supply chain for supply chain management.

Extensible markup language A specification developed by the World Wide Web Consortium, used for defining data elements on a Web page and business-to-business documents. The extensible markup language (XML) uses hypertext markup language (HTML) tag structures. HTML only defines how the elements are displayed; XML defines what the elements contain. This allows the designer to create their own customized tags, enabling the definition, transmission, validation, and interpretation of data between applications and between organizations.

Extranet A data communications network leveraging the Internet as an enabling technology for external enterprise integration.

Integration hub The fundamental c-commerce platform for a standards-based, secure, shared data environment enabled by business process workflow technology for value chain management between trading partners that operate as a virtual enterprise.

Intranet A data communications network leveraging the Internet as an enabling technology for internal enterprise integration.

Private trading exchange An environment that provides real-time interoperability with multiple trading partners, automates and integrates complex business processes across the value chain, secures proprietary information from competitors, and facilitates deployment of agile business models.

Supply chain The community of organizations with the assets, processes, and information that provide elements of supply for a product.

Value chain A trading partner community with mutually beneficial interests spanning from the initial suppliers' supply chain to the final customer's demand chain.

Virtual enterprise integration The use of information technology and collaborative business practices to enable semantic interoperability for the dynamic management of the organizational boundaries in the value chain between trading partners.

CROSS REFERENCES

See *Electronic Commerce and Electronic Business; Extranets; Intranets; Supply Chain Management; Value Chain Analysis.*

REFERENCES

Bauer, M., Poirier, C., Lapide, L., & Bermudez, J. (2001). *E-Business: The strategic impact on supply chain and logistics.* Oak Brook, IL: Council of Logistics Management.

Bond, B., Burdick, D., Miklovic, D., Pond, K., & Eschinger, C. (1999, August 3). *C-commerce: The new arena for business applications* (Research Note). Stamford, CT: Gartner Group.

Bowman, R. (2002, May). Access to data in real time: Seeing isn't everything. *Global logistics & supply chain strategies.* Retrieved July 23, 2002, from http://www.supplychain-brain.com

Bunce, P., Fraser, R., & Hope, J. (2002, June). *Beyond budgeting.* White Paper. Lymington, Hampshire, England: Consortium for Advanced Manufacturing International.

Chandrashekar, A. (1999). Toward the virtual supply chain: The convergence of IT and organization. *The International Journal of Logistics Management, 10*(1), 27–39.

Deloitte Consulting & Deloitte & Touche. (2002). *Directions in collaborative commerce: Managing the extended enterprise.* New York: Deloitte Research. Retrieved March 25, 2003, from http://www.dc.com/asp/pdfviewer.asp?reportname=/pdf/cc_managing.pdf&title=Directions%20in%20Collaborative%20Commerce.

Fingar, P., & Aronica, R. (2001). *The death of "e" and the birth of the real new economy: Business models, technologies and strategies for the 21st century.* Tampa, FL: Meghan-Kiffer Press.

Halpern, M. (2001, February 23). *CPC: Exploiting e-business for product realization* (Strategic Analysis Report). Stamford, CT: Gartner Group.

Hammer, M. (2001). *The agenda: What every business must do to dominate the decade.* New York: Crown Business.

Hoffman, K. (2002, May). Performance data provides key to continuous improvement. *Global logistics & supply chain strategies.* Retrieved July 23, 2002, from http://www.supplychain-brain.com

Kaplan, R., & Norton, D. (2001). *Strategy-focused organization: How balanced scorecard companies thrive in the new business environment.* Boston: Harvard Business School Press, 275–276.

Keebler, J., Manrodt, K., Durtsche, D., & Ledyard, M. (1999). *Keeping score: Measuring the business value of logistics in the supply chain.* Oak Brook: Council of Logistics Management.

Keenan, F., & Ante, S. (2002, February 18). The new teamwork. *Business Week,* EB12–16.

Konicki, S. (2002, July 1). Sears, Michelin Test Supply Chain. *Information Week,* 47–48. Retrieved March 27, 2003, from http://www.informationweek.com/story/IWK20020628S0002 .

Lambert, D., & Pohlen, T. (2001). Supply chain metrics. *The International Journal of Logistics Management, 12*(1), 1–19.

Lee, H. (2000, July/August). Creating value through supply chain integration. *Supply Chain Management Review,* 30–40.

Lee, H. (2001). *Ultimate enterprise value creation using demand-based management* (Stanford Global Supply Chain Management Forum, SGSCMF-W1–2001). Stanford: Stanford University, Graduate School of Business. Retrieved March 26, 2003, from http://www.stanford.edu/group/scforum/

Lee, H., & Whang, S. (2001). *E-Business and supply chain integration* (Stanford Global Supply Chain Management Forum, SGSCMF-W2–2001). Stanford: Stanford University, Graduate School of Business. Retrieved March 26, 2003, from http://www.stanford.edu/group/scforum/

Lee, H., & Whang, S. (2002). The impact of the secondary market on the supply chain. *Management Science, 48*(6), 719–731.

McKenna, R. (1997). *Real time: Preparing for the age of the never satisfied customer.* Boston: Harvard Business School Press.

Moad, J. (2001, December 5). *Amid war, U.S. defense revamps supply chain operations.* Retrieved March 25, 2003, from http://asia.cnet.com/newstech/applications/0,39001103,39004322,00.htm

Murphy, J. (2002, May). Seeing inventory in real time lets you have and not hold. *Global logistics & supply chain strategies.* Retrieved March 27, 2003, from http://www.supplychainbrain.com/archives/5.02.inventory.htm?adcode=10

Sabath, R., & Fontanella, J. (2002, July/August). The unfulfilled promise of supply chain collaboration. *Supply Chain Management Review,* 24–29.

Siegele, L. (2002, February 2). How about now? A survey of the real-time economy. *The Economist.*

Strader, T. J., & Lin, F. R. (1998). Information infrastructure for electronic virtual organization management. *Decision Support Systems, 23*(1), 75–94.

Supply-Chain Operations Reference-Model: SCOR Users Seminar (2002, November 13). From modeling and simulation to real time decision support. Pittsburgh: Supply-Chain Council.

Swanson, S. (2002, July 1). Shopping for savings. *InformationWeek,* 47–48. Retrieved March 27, 2003, from http://www.informationweek.com/story/IWK20020628S0003

Swanson, S. (2002, July 1). Get together. *Information Week,* 37–45. Retrieved March 27, 2003, from http://www.informationweek.com/story/IWK20020627S0010

Voluntary Interindustry Commerce Standards (VICS) Association (2002, July 23). *VICS home page.* Retrieved March 26, 2003 from http://www.cpfr.org

Common Gateway Interface (CGI) Scripts

Stan Kurkovsky, *Columbus State University*

INTRODUCTION

Common gateway interface, CGI for short, is the most common method for passing information from an HTML (hypertext markup language) page in the client's browser to an application running on the Web server, which can process that information (Kew, 2000; CGI, 1999). CGI is completely independent from the browser on the client computer. In fact, it works with any browser, because, unlike HTML content, it does not have to be loaded into and interpreted or executed by the client browser. CGI is not a programming language. As the name explains, it is an interface, or a set of rules that allows an external program to receive an input from a Web browser and produce an output in form of an HTML page.

ORIGINS OF CGI

The Internet is ubiquitous these days, but it is far from being a brand new technology. It has always contained a variety of protocols for exchanging information, but the advent of Web browsers induced Internet's sudden explosive growth. Today, when people hear about the Internet, they associate it with the Web. What makes the Web so attractive that it has become a part of our everyday lives and a business vehicle for many companies? What distinguished the Web apart from other Internet technologies, such as Gopher and FTP (file transfer protocol), was its visual appeal and true interactivity. Even the earliest versions of the HTML standard provided tools for adding images to formatted text, which clearly makes the Web more attractive than a typically command line-based FTP. The Web is interactive and it is not only because the users can browse it by following the links leading from one page to another, but also because the Web has become a true dynamic resource. In the early days of the existence of the World Wide Web, it was merely a collection of personal home pages and academic and corporate Web sites containing a variety of information. All these Web sites were static because they did not change from one request to another. CGI was the first technology that allowed the Web to become dynamic. For example, when the user fills out a registration form, a CGI script would verify that all fields are filled in and contain valid information—there is both first and last names, the telephone number contains a 3-digit area code and a 7-digit number, etc. If the form is incomplete or the data appear to be invalid, the same CGI script would present the user with a message explaining how to correct the information on the form. The same script would add this information into the database and inform the user about the successful registration in case if there were no errors in the data submitted from the form.

Despite the fact that there are many technologies competing with CGI, it remains one of the most common platforms for developing Web applications because of its many advantages. The foremost advantage of using CGI scripts is that this is a true cross-platform technology. CGI scripts work with any Web browser; they also work with most (if not all) Web servers running on Windows and UNIX. The second advantage is that CGI scripts can be written in almost any language the programmer chooses. Such languages as C++ or Perl offer the best support for implementing CGI scripts. The interface of CGI is very simple and does not require any special libraries. For the most part, CGI scripts can rely on the operating system's application programming interface (API) for standard input and output and use the environment variables to communicate with the Web server.

CGI ARCHITECTURE

Usually, when the user clicks on a link in an HTML page or simply types a URL directly in the browser's address field, the browser sends a hypertext transport protocol (HTTP) request to the Web server. If the requested document is a static HTML page, the Web server simply finds the corresponding HTML document on its file system. If the requested HTML document was successfully located on the server file system, it is then returned with the server's HTTP response to the client's browser. However, when a Web server receives a request for a CGI script, the Web server passes some parameters to this script and executes it. The script runs and generates some output, which is then collected by the Web server and returned to the client's browser (Figure 1).

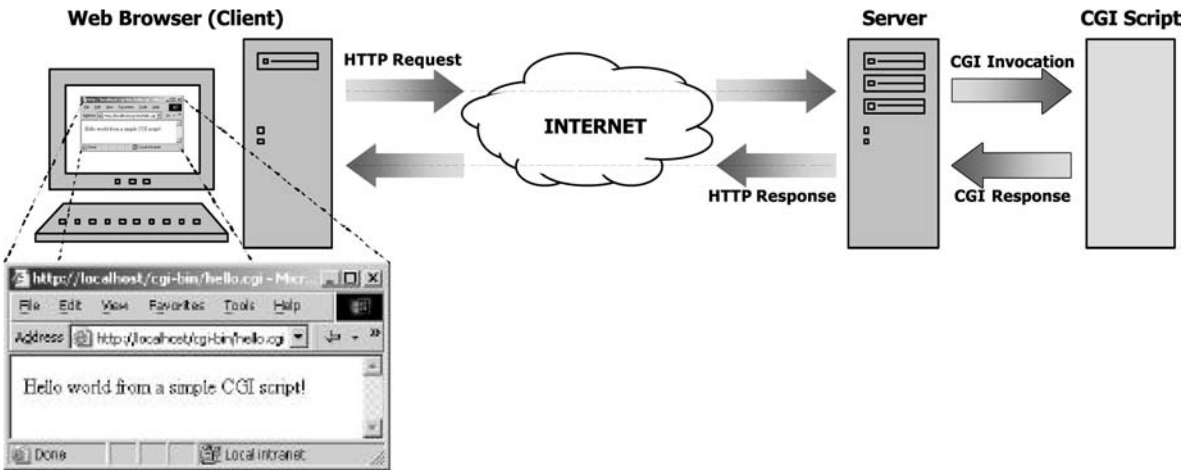

Web Browser (Client)
HTTP Request
INTERNET
Server
CGI Script
CGI Invocation
HTTP Response
CGI Response
http://localhost/cgi-bin/hello.cgi - Micr...
File Edit View Favorites Tools Help
Address http://localhost/cgi-bin/hello.cgi
Hello world from a simple CGI script!
Done
Local intranet

Figure 1: HTTP request for a CGI script.

Let us consider this simple CGI script written in Perl:

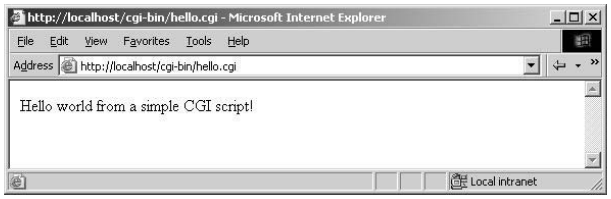
```
#!/usr/bin/perl -w

print <<HELLO_WORLD;
Content-type: text/html

Hello world from a simple CGI script!
HELLO_WORLD
```

Assuming that this script is pointed to by URL http://localhost/cgi-bin/hello.cgi, the user clicks on a link to this URL or simply types this URL in the address field of the browser. These actions will produce the output shown in Figure 2 in Microsoft Internet Explorer 6.

A CGI script has dynamically generated the HTML content, which is displayed in the browser. Before the user could see such an output, the user's Web browser had to send an HTTP request to the Web server. CGI scripts are requested in very much the same way as regular HTML documents—their address is passed to the browser in a URL. Figure 3 presents some common components of a URL that can be used to request a CGI script.

Typically, CGI scripts are stored in a special directory on the Web server's file system, which is mapped into the virtual directory named cgi-bin. Also, CGI scripts commonly have a specific extension, such as .cgi. Data from an HTML page can be passed to a CGI script in the query part of the URL by using an HTTP GET request. CGI scripts can also use an HTTP POST request to receive data from HTML forms. This is discussed later.

The URL used in our first example consists of the components shown in Figure 4.


http://localhost/cgi-bin/hello.cgi - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://localhost/cgi-bin/hello.cgi
Local intranet

Hello world from a simple CGI script!

Figure 2: Output of a simple CGI script.

Requesting such a URL by clicking on a corresponding link would generate the following HTTP GET request:

```
GET /cgi-bin/hello.cgi HTTP/1.1
Host: localhost
```

When the Web server receives such a request, it typically checks the file extension of the requested resource (/cgi-bin/hello.cgi in this case) to determine the type of the document. If the .cgi extension is associated with a CGI script on the Web server and it recognizes the /cgi-bin as the directory for CGI scripts, the Web server will execute script hello.cgi instead of treating it as a static HTML document.

When a CGI script's process is invoked, it exists in an environment created for it by the Web server. Typically, this environment includes certain predefined variables and standard file handles. Specifics of this environment may sometimes vary between different Web servers. Table 1 lists some standard CGI environment variables (*CGI environment variables*, n.d.).

A simple Perl script can be used to produce an alphabetic list of all variables currently defined in the environment of a Web server:

```
#!/usr/bin/perl -w

use strict;

print "Content-type: text/html\n\n";

print "<TABLE>";

my $EnvVarName;
foreach $EnvVarName (sort keys %ENV) {
  print "<TR><TD><B>$EnvVarName</B></TD>";
  print "<TD>$ENV{$EnvVarName}</TD></TR>\n";
}

print "</TABLE>"
```

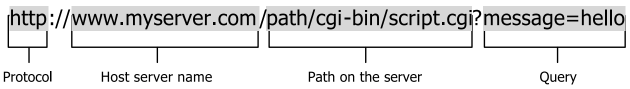

http://www.myserver.com/path/cgi-bin/script.cgi?message=hello
Protocol
Host server name
Path on the server
Query

Figure 3: Structure of a URL.

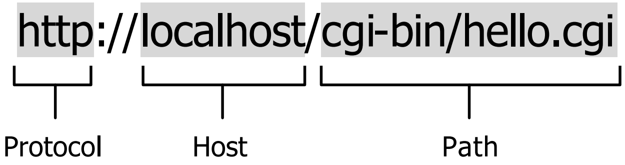


Figure 4: Structure of a simplified URL.

Running the script above on Windows 2000 with Internet Information Services 5.0 produces the output, a fragment of which is shown in Figure 5. Note that not all the standard environment variables are present and several other variables are added by the server.

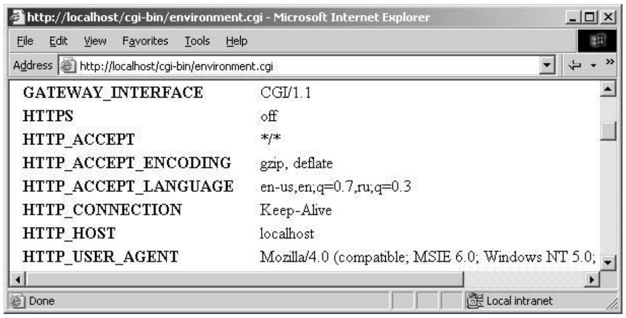
Beside the environment variables, a Web server creates several standard file handles that are available to the CGI script:

STDIN contains encoded data from an HTML form when it is submitted using POST request. The length of valid data in STDIN is determined by the value of the *content-length* HTTP header. For GET requests, STDIN is empty.

STDOUT is used by CGI scripts to create their HTML output. It may include some HTTP headers.

STDERR output typically is used for logging errors that occurred in a CGI script. However, particular details of handling STDERR output depend on the specifics of the HTTP server.

When a CGI script generates an HTML output, it must be preceded by some HTTP headers to be successfully interpreted by the Web server and then delivered to the client browser. CGI scripts do not need to generate complete HTTP headers—the Web server typically will complete the header as long as one of the following partial headers are present:


http://localhost/cgi-bin/environment.cgi - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://localhost/cgi-bin/environment.cgi
Done
Local intranet

```
GATEWAY_INTERFACE      CGI/1.1
HTTPS                  off
HTTP_ACCEPT            */*
HTTP_ACCEPT_ENCODING   gzip, deflate
HTTP_ACCEPT_LANGUAGE   en-us,en;q=0.7,ru;q=0.3
HTTP_CONNECTION        Keep-Alive
HTTP_HOST              localhost
HTTP_USER_AGENT        Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.0;
```

Figure 5: Output of a CGI script listing standard environment variables.

Content-type specifies the media type that will be produced by the CGI script;

Location specifies a URL and is used for redirecting the output; and

Status header is used to exchange information between the CGI script and the Web server.

In order for the Web server to successfully create an HTTP response and deliver the resulting document to the client browser, the CGI script must indicate the media type that the document contains. In most cases, when a CGI script generates an HTML document, it must include the following content-type header:

```
Content-type: text/html
```

The content-type header may contain any valid media type. For example, the following header indicates that the

Table 1 Some Standard CGI Environment Variables

AUTH_TYPE	If the server supports user authentication, and the script is protected, this is the protocol-specific authentication method used to validate the user.
CONTENT_LENGTH	The length of the said content as given by the client.
CONTENT_TYPE	For queries that have attached information, such as HTTP POST and GET, this is the content type of the data.
GATEWAY_INTERFACE	The revision of the CGI specification to which this server complies.
HTTP_ACCEPT	The MIME types which the client will accept, as given by HTTP headers.
HTTP_USER_AGENT	The browser the client is using to send the request.
PATH_INFO	The extra path information, as given by the client.
PATH_TRANSLATED	The server provides a translated version of PATH_INFO, which takes the path and does any virtual-to-physical mapping to it.
QUERY_STRING	The information which follows the ? in the URL which referenced this script.
REMOTE_ADDR	The IP address of the remote host making the request.
REMOTE_HOST	The hostname making the request.
REQUEST_METHOD	The method with which the request was made.
SCRIPT_NAME	A virtual path to the script being executed, used for self-referencing URLs.
SERVER_NAME	The server's hostname, DNS alias, or IP address as it would appear in self-referencing URLs.
SERVER_PORT	The port number to which the request was sent.
SERVER_PROTOCOL	The name and revision of the information protocol of the request.
SERVER_SOFTWARE	The name and version of the information server software answering the request.

CGI script returns a single JPEG image:

```
Content-type: image/jpeg
```

It is imperative that the content-type header is followed by a blank line. This tells the Web server that all subsequent output generated by a CGI script should be treated as content, but not as a part of HTTP header. The simple Perl script below correctly creates a content-type header that indicates that the CGI script's output contains an XML document:

```
#!/usr/bin/perl -w

print "Content-type: text/xml\n\n";
```

Quite often a CGI script may serve as a sort of dispatcher that, instead of generating customized output, simply returns different URLs upon examining the input. CGI scripts can forward the client browser to a specific URL by using the location header. The following Perl script generates an HTTP header that redirects the client browser to an HTML page mypage.html.

```
#!/usr/bin/perl -w

print "Location: mypage.html\n\n";
```

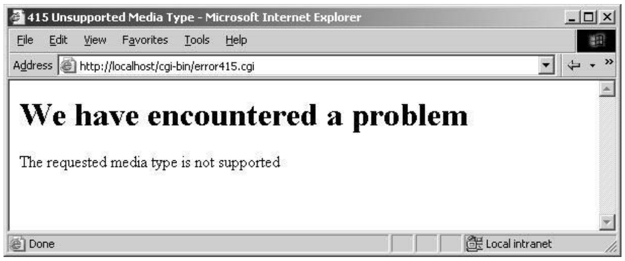
The location header may contain absolute as well as relative URLs. A relative URL containing a full path is resolved by the Web server, which opens the referenced resource and returns it to the client browser as if it were the output of the CGI script. In all other cases (if it is an absolute URL or a relative URL containing a relative path), the generated URL is sent to the client browser. Now it is a responsibility of the client browser to generate an HTTP request for the resource referenced in the received URL. Most current Web browsers do this automatically.

The status header does not directly map into an HTTP header—it is only used to exchange information between the CGI script and the Web server. If the execution of a CGI script was successful, there is no need to specify the status header. In such a case the Web server will automatically add the following HTTP status headers:

200 OK if the CGI script returned a content-type header, or

302 Found if the CGI script returned a location header.

If the execution of a CGI script was unsuccessful, the status header can be used as a tool to communicate the error code and error message back to the Web server. It is important to use a standard status code (Fielding et al., n.d.). Such a code should be chosen to fit the actual error that occurred in a CGI script. For example, HTTP status code *415 Unsupported Media Type* can be generated by a CGI script, which does not know how to handle a particular media type. Whenever a CGI script chooses to return an error status code, it should also return a content-type header and a user-friendly HTML document describing the nature of the occurred problem in less technical terms.


415 Unsupported Media Type - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address
http://localhost/cgi-bin/error415.cgi
Done
Local intranet

We have encountered a problem

The requested media type is not supported

Figure 6: Typical use of HTTP status header.

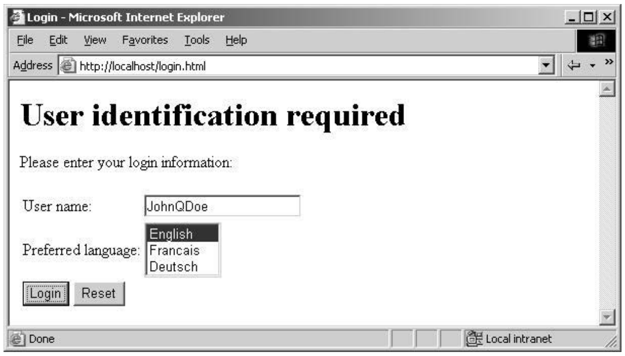
For example, the following Perl script generates such a message:

```
#!/usr/bin/perl

print <<END_OF_ERROR_MSG;
Status: 415 Unsupported Media Type
Content-type: text/html

<HTML>
   <HEAD>
      <TITLE>415 Unsupported Media Type
        </TITLE>
   </HEAD>
   <BODY>
      <H1>We have encountered a problem
        </H1>
          <P>The requested media type is
            not supported</P>
   </BODY>
</HTML>
END_OF_ERROR_MSG
```

The CGI script above generates the output shown in Figure 6.

Approaching this problem from a more practical perspective, nowadays most CGI script developers choose not to use a status header. They simply prefer to write scripts that always return a content-type header (with a success status code added by the Web server by default). If an error occurs they handle it by the logic embedded in their scripts and communicate the nature of the problem to the user by including a corresponding error message in the generated HTML document.

The most common way to invoke a CGI script is by submitting input from an HTML form. Forms provide a very flexible tool for building front-ends for all Web-enabled applications, including those implemented with CGI scripts. In most cases, HTML forms are used for two related purposes: collecting input and accepting commands. These tasks are often related because frequently the nature of a command depends on the context specified in the user input. In a typical interactive Web application, HTML forms can be used to collect data, which is then stored in a database located on the Web server. For example, such an application may maintain a database of its users, who can log in to the Web site to access its personalized content. Figure 7 shows a typical example of a user login form needed to enter the Web application described above.


Login - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://localhost/login.html
Login
Reset
Done
Local intranet

User identification required

Please enter your login information:

User name: JohnQDoe

Preferred language: English Francais Deutsch

Figure 7: CGI script output—user identification form.

Such a form can be generated by the HTML code.

```
<HTML>
<HEAD>
<TITLE>Login</TITLE>
</HEAD>

<BODY>

<H1>User identification required</H1>

<P>Please enter your login information:</P>

<FORM NAME="LOGIN" ACTION="/cgi-bin/
  login.cgi" METHOD="POST">
<TABLE BORDER="0">
  <TR><TD>User name:</TD>
      <TD><INPUT TYPE="TEXT" NAME=
        "USERNAME"></TD></TR>
  <TR><TD>Preferred language:</TD>
      <TD><SELECT NAME="LANGUAGE" SIZE=3>
          <OPTION SELECTED>English</OPTION>
          <OPTION>Francais</OPTION>
          <OPTION>Deutsch</OPTION>
          </SELECT></TD></TR>
  <TR><TD COLSPAN=2><INPUT TYPE="SUBMIT"
    VALUE="Login">
                  <INPUT TYPE="RESET"
                    VALUE="Reset"></TD>
                    </TR>
</TABLE>
</FORM>

</BODY>
</HTML>
```

It is important to notice the names of the HTML form tags because they are used to pass the entered information to the Web server and then to a CGI script. Clicking the “Login” button on this form would generate the HTTP request

```
POST/cgi-bin/login.cgi HTTP/1.1
Host: localhost
Content-Length: 34
Content-Type: application/x-www-form-
  urlencoded

USERNAME=JohnQDoe&LANGUAGE=English
```

When this request is passed by the Web server to a CGI script, because this is a POST request, the Web server removes all HTTP headers and passes the remainder of the message to the CGI script:

```
USERNAME=JohnQDoe&LANGUAGE=English
```

The difference between POST and GET method, as far as CGI scripts are concerned, lies in the method used for passing on the form data. When a POST request is used, form data are passed through STDIN and the script should read from there. The number of bytes to be read is specified in the content-length header. When a GET request is used, the data are passed in the environment variable QUERY_STRING. The value of the content-type header (typically, application/x-www-form-urlencoded) is the same for both GET and POST requests. If the form shown above were changed to use the GET request, clicking the Login button would generate the HTTP request

```
GET /cgi-bin/login.cgi?USERNAME=JohnQDoe
  &LANGUAGE=English HTTP/1.1
Host: localhost
Content-Type: application/x-www-form-
  urlencoded
```

To process either one of HTTP requests shown above, one might write a CGI script in Perl, C, or any other suitable language. However, processing input generated by HTML forms requires a number of routine manipulations that are reused from one script to another. Perl provides a specialized module, CGI.pm, which is a de facto standard for creating CGI scripts with Perl. Besides parsing the input generated by HTML forms, it also contains routines for generating HTML code fragments including headers and forms. A CGI script taking advantage of CGI.pm module should include the following statement in its beginning:

```
use CGI;
```

A simple CGI script taking advantage of features offered by CGI.pm module can be used to process the input generated by the log-in page presented above. All that needs to be done is to generate appropriate HTTP headers and produce a customized HTML output based on the user selection:

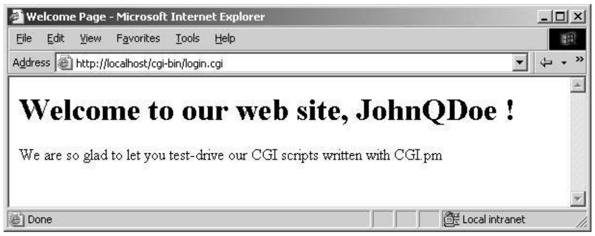
```
#!/usr/bin/perl -w

use strict;
use CGI;

my $cgi = new CGI;
my $message;

print $cgi->header("text/html"),
      $cgi->start_html(-title => "Welcome
        Page");

if ($cgi->param("LANGUAGE") eq "English") {
     $message = "Welcome to our web site,";
}
```

```
elsif ($cgi->param("LANGUAGE") eq
  "Francais") {
      $message = "Bienvenue a notre site
        web,";
}
elsif ($cgi->param("LANGUAGE") eq
  "Deutsch") {
      $message = "Willkommen zu unserer web
        site,";
}
else {
      $message = "We do not know what
        language you speak,";
}
print $cgi->h1($message, $cgi->param
  ("USERNAME"), "!"),
       $cgi->p("We are so glad to let you
         test-drive our CGI scripts written
         with CGI.pm"),
       $cgi->end_html;
```

To access any feature offered in CGI.pm, the script creates a corresponding object called cgi. It includes collection param, which contains parsed parameters received by the Web server from the client and passed into the script. This script also takes advantage of HTML-generating capabilities of CGI.pm and creates the following HTML output (indentations and line breaks have been added):

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE html
   PUBLIC "-//W3C//DTD XHTML Basic 1.0//EN"
   "http://www.w3.org/TR/xhtml-basic/
     xhtml-basic10.dtd">
<html xmlns="http://www.w3.org/1999/xhtml"
  lang="en-US">
   <head>
      <title>Welcome Page</title>
   </head>
   <body>
      <h1>Welcome to our web site,
        JohnQDoe !</h1>
      <p>We are so glad to let you
        test-drive our CGI scripts written
        with CGI.pm</p>
   </body>
</html>
```

Figure 8 shows what this HTML code looks like when displayed in Internet Explorer 6.


Welcome Page - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address
Done
Local intranet

http://localhost/cgi-bin/login.cgi

Welcome to our web site, JohnQDoe !

We are so glad to let you test-drive our CGI scripts written with CGI.pm

Figure 8: CGI script output—greetings.

A major challenge of any CGI application is the statelessness of HTTP. Most interactive Web applications must identify the sequence of page requests made by the same user. It is difficult to underestimate the importance of maintaining state and collecting information about each individual user's session (series of interactions with the Web site). For example, many travel Web sites allow their users to build flexible itineraries. In many cases, the user needs to enter some data on every page in a long sequence of pages (enter departure and return date and location, then select the departing flight and then select the return flight). Building a complex itinerary may require entering some information or clicking on buttons on every one of five or more pages. Such a travel Web site must track each user through their individual sequences of visited pages and collect all entered data to assemble a complete itinerary. There are several different ways to maintain state in a CGI application. These methods include using query strings to pass additional parameters, using hidden form fields, and manipulating client-side cookies.

A CGI script behind a large-scale Web application typically would take advantage of many (if not all) HTML form tags. Also, it is typical for the same CGI script not only to process the user input from an HTML form, but also to generate the form itself. Such an approach is illustrated below by a demonstration of a pizza-ordering application implemented as a single CGI script. This is a typical scenario of the user interaction with the CGI script of this application:

The CGI script is reached by an ordinary link or its URL is typed directly in the address field of the browser

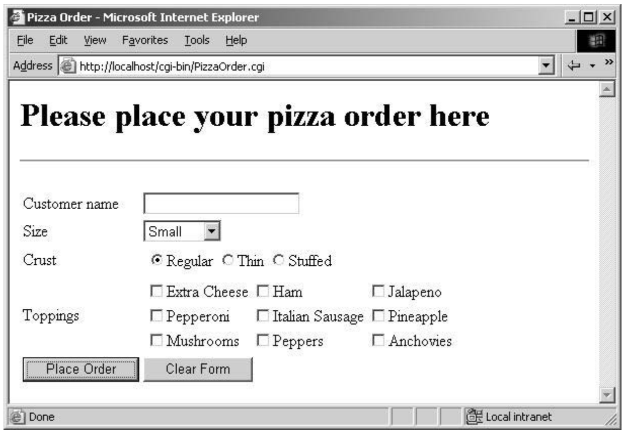
Since the CGI script sees no form data, it generates the form.

The user fills this out, and submits it, and the data are sent back to the same CGI script.

Since the CGI script sees that it has been sent data, it now generates the confirmation message.

The following example uses a single CGI script that generates and processes customized pizza orders:

```
#!/usr/bin/perl -w

use strict;
use CGI;

my $cgi = new CGI;
my $error_message;

print $cgi->header("text/html"),
      $cgi->start_html( -title =>
        "Pizza Order");

if(defined ($cgi->param("customer"))) {
   if($cgi->param("customer") eq "") {
      $error_message = $cgi->p($cgi->b
        ("Order incomplete - please enter
        customer name"));
      print_form();
   }
   else {
      print_order();
   }
}
```


Pizza Order - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://localhost/cgi-bin/PizzaOrder.cgi
Done
Local intranet

Please place your pizza order here

Customer name □

Size Small

Crust ● Regular ○ Thin ○ Stuffed

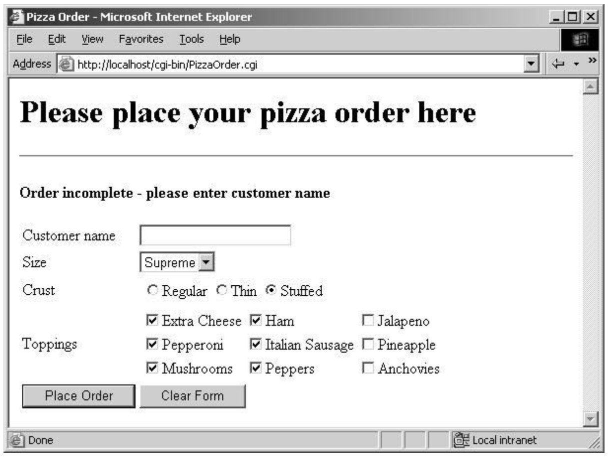
Toppings □ Extra Cheese □ Ham □ Jalapeno □ Pepperoni □ Italian Sausage □ Pineapple □ Mushrooms □ Peppers □ Anchovies

Place Order | Clear Form

Figure 9: CGI script output—pizza order form.

```
else {
   print_form();
}

print $cgi->end_html;
```

To determine whether the script currently needs to generate the input form or process the user input from this form, this CGI script checks whether parameter "customer" is defined. CGI.pm generates it upon examining the HTTP GET or POST request passed into the CGI script. The data in this parameter are created by the user input in the form, particularly, by filling out the field "Customer name." If this parameter is not defined, it means that this CGI script was invoked directly and it needs to generate the form. The Perl sub procedure print_form generates the output shown in Figure 9:


Pizza Order - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://localhost/cgi-bin/PizzaOrder.cgi
Done
Local intranet

Please place your pizza order here

Order incomplete - please enter customer name

Customer name □

Size Supreme

Crust ○ Regular ○ Thin ● Stuffed

Toppings ☑ Extra Cheese ☑ Ham □ Jalapeno ☑ Pepperoni ☑ Italian Sausage □ Pineapple ☑ Mushrooms ☑ Peppers □ Anchovies

Place Order | Clear Form

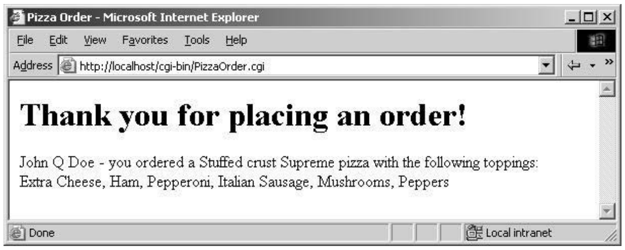
Figure 10: CGI script output—errors on the form.

Note that the $cgi->start_form call generates a form tag that specifies the current script as the form action. Variable $error_message may contain any error messages that will be included in the output if the submitted information is incomplete. In particular, this script checks whether the customer's name was specified. If it was not, the script will not process the order and display the form once again with the error message indicating any input discrepancies (Figure 10).

Once the order form is complete, this CGI script generates an order confirmation by calling sub procedure print_order. This sub procedure generates the HTML output presented in Figure 11:

```
sub print_order {
   my @toppings_list = $cgi->param
     ("toppings");
```

```
sub print_form {
  print $cgi->h1("Please place your pizza order here"),
     $cgi->hr,
     $error_message,
     $cgi->start_form,
     $cgi->table( { -border => 0},
        $cgi->Tr( [
           $cgi->td( ["Customer name", $cgi->textfield(-name => "customer")]),
           $cgi->td( ["Size," $cgi->popup_menu(-name => "size",
                      -values=> ["Small", "Medium", "Large", "Supreme"]) ]),
           $cgi->td( ["Crust", $cgi->radio_group(-name => "crust",
                      -multiple =>0, -columns=>3,
                      -values => ["Regular", "Thin", "Stuffed"]) ]),
           $cgi->td( ["Toppings", $cgi->checkbox_group(-name =>"toppings",
                      -columns=>3,
                      -values => ["Extra Cheese", "Pepperoni", "Mushrooms",
                      "Ham", "Italian Sausage", "Peppers",
                      "Jalapeno," "Pineapple," "Anchovies"]) ]),
           $cgi->td( [$cgi->submit(-value =>" Place Order "),
                     $cgi->reset(-value =>" Clear Form ") ]),
        ])
     ),
  $cgi->end_form;
}
```


Pizza Order - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://localhost/cgi-bin/PizzaOrder.cgi
Done
Local intranet

Thank you for placing an order!

John Q Doe - you ordered a Stuffed crust Supreme pizza with the following toppings:
Extra Cheese, Ham, Pepperoni, Italian Sausage, Mushrooms, Peppers

Figure 11: CGI script output—order confirmation.

```
    print $cgi->h1("Thank you for placing
      an order!"),
           $cgi->p($cgi->param("customer"),
             " - you ordered a ",
           $cgi->param("crust"), "crust",
           $cgi->param("size")," pizza with ",
           (@toppings_list ? ("the following
               toppings:",
                          $cgi->br, join(", ",
                            @toppings_list)) :
"no toppings"));
}
```

Such an iterative approach is used in most CGI scripts designed for processing user input. First, the script generates the input form. Second, the same script is used to verify the completeness and correctness of the entered data. Third, the data are processed by the script and typically saved in a server database. Note that instead of verifying the completeness of the data entered in the form with the help of the CGI script on the server side, it is often more convenient to use JavaScript on the client side and make such verifications before the form is submitted to the Web server. Once the data from the form have been accepted and analyzed, CGI script either generates the confirmation output or moves on to presenting the user with the next form. User input in multiple forms may be necessary in large e-commerce applications when there is a need to collect a large amount of information. For example, in an online bookstore, it may be necessary to enter customer information in several steps:

Enter mailing address
Enter payment information
Verify shopping basket content and order amount
Confirm order

CGI LANGUAGES AND ALTERNATIVES TO CGI

Most examples in this article are written in Perl. However, almost any programming language can be used for writing CGI scripts. A programming language must be able to produce an executable program that would satisfy the following two criteria in order to be used for CGI scripting:

The program must be text-based and not require a GUI, and

The program must generate a valid HTML content-type header.

The following is a list of some common programming languages that can be used for writing CGI scripts.

Perl

Currently, Perl is the dominant language for writing CGI scripts (Wall et al., 2000; Castro, 2001). This is because Perl is easy to use and to learn (it is similar to C), it has all the features needed for CGI programming (transparent IO and access to environment variables), and it is a very popular scripting language for UNIX—the first platform to support Web servers. Another factor contributing to the popularity of Perl is a wide availability of libraries that facilitate writing CGI scripts. The CGI.pm module makes a number of CGI-related tasks easier (in particular, it converts the data from form input into a data structure) and usually it is bundled with Perl interpreters. Because Perl was the first widely used CGI programming language, there is a lot of open source CGI scripts that have been written over the years and are freely available on the Internet. Very often it is very easy to download a working script and adapt it to the particular problem at hand. From the programming point of view, Perl is appealing because it contains extremely powerful string manipulation operators, it easily interfaces with external applications, and in cases when errors occur, Perl generates specific and detailed error messages (Guelich, Gundavaram, & Birznieks, 2000). The only possible disadvantage of using Perl for CGI scripting is that it is an interpreted language and therefore its performance may be diminished. However, most implementations of Perl compile source code into low-level intermediate code and then execute them.

C/C++/C#

Just like any other compiled languages capable of creating command-line executables, C and C++ are good candidates as programming languages for writing CGI scripts (Deitel & Deitel, 2002). Compared to the fastest Perl script, most likely a similar C/C++ program will start and execute faster because it has already been compiled, while a Perl script will have to be interpreted before it can be executed.

It is important to compare scripting languages (such as Perl) to compiled languages (such as C/C++) from the perspective of using them for CGI scripting. The main advantages of compiled languages are:

Performance. Programs written in a compiled language are always readily available to execute and do not incur an overhead of loading an interpreter every time they need to run. Performance is crucial in the Web environment when a CGI script may have to process numerous requests in a short period of time.

Flexibility. Most compiled languages are general-purpose languages. That is, they are designed to suit almost any programming task. Scripting languages, in contrast, almost always are designed for a particular type of task. In some cases, scripting languages may lack a particular feature needed to implement a particular CGI script.

The main advantages of scripting languages include:

Rapid development. Most scripting languages have higher-level statements than those of common-purpose compiled languages. For example, Perl has a lot of tools specifically designed to handle strings, regular expressions, and text files. Using C or C++ to write a CGI application with these features would require more lines of code than that of a similar Perl program.

Learning curve. In general, scripting languages are easier to learn than compiled languages. As a rule, scripting languages are more forgiving to a novice programmer and do not require the knowledge of all control structures available in the language to efficiently implement a given task.

Debugging. Given their interpreted nature, programs written in scripting languages do not need to be recompiled each time a change is made during debugging. To some extent, this may speed-up the process of debugging.

Despite the fact that C# belongs to the same family of programming languages as C and C++, C# is not very well suited for writing CGI scripts for a number of reasons. There are C and C++ compilers for virtually any operating system and it makes C/C++ code portable. C# programs can only be compiled and executed on the Microsoft .NET framework. C# programs are compiled into CLR (common language runtime), which caries a typical interpreting overhead, which may not be desired in a time-critical Web application. On the other head, C# and Virtual Basic .NET are perfectly suited for creating ASP.NET applications, which are discussed below.

Java

Java programs must be executed using a Java Virtual Machine and, given its architecture, it is impossible to call a CGI script written in Java directly from a Web server. It is possible, however, to create a simple wrapper program in Perl or C, which would invoke the Java program. This adds a double overhead of first calling a wrapper program and then starting the Java Virtual Machine, which would then run the Java CGI script.

Nevertheless, there are a number of options available for applying Java for creating interactive Web applications, which are discussed below.

CGI has started a revolution intended to change the nature of the Web and make it an interactive medium. It would not be wise to assume that such a good idea could be left without being explored further. CGI has spawned a number of technologies that build upon the same idea—accept and respond to queries from Web browsers and present dynamic content using HTTP protocol. Most of them address the main problem of CGI architecture—creating a separate process on the Web server every time the script is executed. Some of these technologies attempt to remove the distinction between the script and the HTML page by mixing them together.

ASP and ASP.NET

Active Server Pages (ASP) were created by Microsoft and served as a scripting language for the Microsoft Internet Information Server. The main advantage of ASP is that the scripting engine is integrated into the Web server and, therefore, it does not require an additional process to run. ASP code can be mixed together with HTML, which eliminates the need to write separate programs. The most popular language for writing ASP scripts is VBScript, essentially a stripped-down version of Visual Basic. ASP also supports JScript—Microsoft's version of JavaScript; other languages are also available. Initially, ASP was only available on Windows-based servers, but now it has been adapted for other platforms as well.

ASP.NET is a new version of ASP based on the Microsoft.NET framework, which currently supports Visual Basic.NET and C# as scripting languages. ASP.NET also makes a clearer distinction between HTML and code—it somewhat departs from ASP's philosophy of intertwining HTML with scripting statements. Instead, ASP.NET separates the interface of the Web form from its event-driven code, which can be written on any .NET programming language. As of October 2002, the following languages were being adapted to run on the .NET Framework: COBOL, Fortran, Java (J#), Pascal, Perl, and many others.

Java Servlets and JSP

Java servlets follow the same basic idea as CGI scripts in that they run as separate programs that receive requests and generate HTML documents. Java servlets can be thought of as Java applets that run on the server and do not have any graphical user interface. Java servlets are server- and platform-independent and they have full access to the entire family of Java APIs. Java Server Pages (JSP) is an extension to Java servlet technology, and it is very similar to ASP in that it allows developers to embed Java code into HTML pages.

PHP

PHP (PHP: Hypertext Processor) is an open source HTML-embedded scripting language similar to Perland functions in the same fashion as ASP. It was initially developed to run on the Apache server, but has been adapted to run on other Web servers as well.

CGI APPLICATIONS AND THE FUTURE OF CGI

Despite the fact that CGI may be considered an aging technology and the fact that there are many competing technologies, CGI scripts are still very widely used in the Internet and there is no reason to suggest any decline in its popularity. Many Web sites contain directories of freely available CGI scripts that developers can use in their applications (CGI Resource Index, n.d.; HotScripts, n.d.). An extensive list of CGI script directories can be found in the *Google CGI Web Directory* (n.d.). There are several reasons why CGI scripts still remain a viable tool for Web application development:

CGI scripts are platform independent;

CGI is the oldest technology for enabling Web interactivity, and over the years developers have accumulated many CGI scripts that are now available for free; and

CGI scripts can be written in many languages, so developers do not have to learn a new language.

By far, the most popular class of CGI application is the shopping cart. As of October 2002, HotScripts.com was listing over 100 CGI scripts in Perl for creating shopping carts. Overall, the e-commerce category at HotScripts.com was the second most populated, containing about 170 Perl scripts, only trailing behind the general form processors category with 190 entries. Other e-commerce applications, for which there exist Perl CGI scripts, include:

Billing systems;
Credit card processing;
Currency exchange;
Ordering systems; and
PayPal payment processing.

HotScripts.com contains an exhaustive list of CGI scripts categorized by their application and programming language (ASP, C/C++, JavaScript, Perl, PHP, etc). Among many others, this list contains such script categories as:

Advertisement management;
Calendars;
Chat script;
Content management;
Counters;
Database tools;
Discussion boards;
E-commerce;
E-mail systems;
File manipulation;
Guestbooks;
Image galleries;
Mailing list management;
News publishing;
Password protection;
Polls and voting;
Searching;
Security systems;
Tests and quizzes;
Virtual communities;
Web fetching; and
Web traffic analysis.

Middleware products, such as ASP, JSP, and PHP, may seem to push CGI scripts out of the market. Their primary advantage is that they offer performance advantages—they are coupled with the Web servers and execute in-process. Most CGI scripts execute out-of-process; that is, they require launching a separate program (e.g., a Perl interpreter) to execute. However, despite its relative age and disadvantages, such features of CGI as simplicity, extensibility, portability, transparency, and ease of use make it a strong competitor to these new technologies.

CONCLUSION

The common gateway interface has literally revolutionized the way we use the Internet. CGI scripts were the first technology to enable interactivity on the Web. Without CGI scripts and all other technologies spawned by it, we would still be using static Web pages instead of interactive Web applications. CGI scripts provide the mechanism for the Web server to respond to the client requests by generating Web pages with customized and personalized content. The common gateway interface is platform-independent and transparent to the client browsers. CGI scripts can be written in many programming languages capable of producing command-line executables that can generate HTML output. Historically, Perl, “the duct tape of the Internet,” has been the language most widely used for creating CGI scripts. All other technologies for creating interactive Web applications, such as ASP, JSP, and ColdFusion, were born because of what CGI started—the need for making the Interned a true interactive medium. CGI is older than these competing technologies, which may put it at a disadvantage. However, CGI developers benefit from the age of CGI as there are a countless number of CGI scripts available on the Internet. The benefits of using CGI clearly outweigh its disadvantages and because of that, CGI is here to stay.

GLOSSARY

Active Server Pages (ASP) A scripting environment built into Microsoft Internet Information Server, which allows combining HTML with statements in a scripting language, typically VBScript.

Application programming interface (API) Calling conventions used by applications to access services of the operating system or other applications.

Common gateway interface (CGI) A standard for running external programs on a Web server.

Hypertext markup language (HTML) Instructs software on how to render a Web page.

Hypertext transport protocol (HTTP) Stateless protocol that coordinates transport of HTML documents on the Internet.

Java Server Pages (JSP) An extension of Java Servlet API to generate dynamic Web pages.

PHP: Hypertext Processor (PHP) A cross-platform open-source scripting environment similar to ASP.

CROSS REFERENCES

See *Active Server Pages; C/C++; DHTML (Dynamic HyperText Markup Language); HTML/XHTML (HyperText Markup Language/Extensible HyperText Markup Language); Java Server Pages (JSP); TCP/IP Suite; Visual Basic Scripting Edition (VBScript).*

REFERENCES

Castro, E. (2001). *Perl and CGI for the World Wide Web*. Peachpit Press, Berkeley, CA.

CGI: Common gateway interface (1999). Retrieved from http://www.w3.org/CGI

CGI environment variables (n.d.). Retrieved from http://hoohoo.ncsa.uiuc.edu/cgi/env.html

CGI Resource Index (n.d.). Retrieved from http://www.cgi-resources.com

Deitel, H. M., & Deitel, P. J. (2002). C++: *How to program* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Fielding, et al. (n.d.). *HTTP 1.1: 10 status codes definitions*. Retrieved from http://www.w3.org/Protocols/rfc2616/rfc2616-sec10.html

Google CGI Web Directory (n.d.). Retrieved from http://directory.google.com/Top/Computers/Programming/Internet/CGI/Scripts_and_Programs/Collections_and_Directories

Guelich, S., Gundavaram, S., & Birznieks, G. (2000). *CGI programming with Perl* (2nd ed.). Sebastopol, CA: O'Reilly.

HotScripts (n.d.). Retrieved from http://hotscripts.com

Kew, N. (2000). *CGI programming FAQ*. Retrieved from http://www.htmlhelp.com/faq/cgifaq.html

Wall, L., Christiansen, T., & Orwant, J. (2000). *Programming Perl* (3rd ed.). Sebastopol, CA: O'Reilly.

FURTHER READING

Christenberry, J. (2000). *CGI fast and easy Web development*. Premier Press.

Colburn, R. (2002). *SAMs teach yourself CGI in 24 hours* (2nd ed.). Pearson Education.

Meltzer, K., & Michalski, B. (2001). *Writing CGI Applications with Perl*. Pearson Education.

Stein, L. (1998). *Official guide to programming with CGI.pm*. New York: Wiley.

Torkington, N., & Christiansen, T. (1998). *Perl cookbook*. Sebastopol, CA: O'Reilly.

Computer Literacy

Hossein Bidgoli, *California State University, Bakersfield*

INTRODUCTION

This chapter provides a general computer literacy for readers with little or no computer background. It discusses different data processing systems that were used before the computer era and highlights general capabilities of computers, their unique power, and their applications. The chapter describes the characteristics of successive generations of computers and examines the input-process-output cycle. The chapter explains the differences between data and information and the types of processing including batch and transaction processing. General applications of computers are introduced and basic productivity tools are highlighted. The chapter concludes with a series of guidelines for proper maintenance of a microcomputer system.

WHAT IS COMPUTER LITERACY?

In today's competitive environment, computer literacy is as important as reading and writing literacy. The number of computers in use is increasing on a daily basis throughout the world. Significant price reduction and impressive improvement in their processing power have made microcomputers a household item. There are many different definitions of computer literacy. In this chapter, we define computer literacy as the basic knowledge needed for understanding and using computers for day-to-day tasks. In this regard, the user should have a basic understanding of hardware, software, and the commonly used applications of computers. An increasing number of people are using their microcomputers to connect to the Internet for accessing the wealth of information available in public and private databases. To do this, a basic knowledge of the Web and Web technologies is also needed. Another chapter in this encyclopedia reviews basic Web literacy and instructions for effective utilization of the Internet and its popular applications. In the next section, a brief overview of Internet literacy will be presented. For a comprehensive understanding of computer operations and productivity tools, the references introduced at the end of the chapter should be consulted.

WHAT IS INTERNET LITERACY?

The Internet is a collection of millions of computers and network systems of all sizes. No one actually owns or runs the Internet. Each network is locally administered and funded, in some cases by volunteers. It is estimated that more than 200 countries are directly or indirectly connected to the Internet. This number is increasing on a daily basis and makes the Internet a true companion for both home and office (Bidgoli, 2002).

The Internet started in 1969 as a Defense Department Advanced Research Projects Agency project called ARPANET. It served from 1969 through 1990 as the basis for early networking research, and as a central backbone network during development of the Internet. To use the Internet, a user needs a computer with a modem and an Internet service provider (ISP). As soon as a user is connected, a wealth of information is at his or her fingertips.

Usually most computers are equipped with a graphical browser, such as Microsoft Internet Explorer or Netscape Navigator. An Internet address, or URL (universal resourse locater), is entered through the browser software, and by pressing the Enter key on the keyboard or clicking the mouse, the user is connected to a site. The Internet site could be across the hall or on a different continent! As soon as the URL is entered, the browser software searches the Web and provides the user with the requested information. Throughout this encyclopedia, a wealth of information and many applications that are readily available on the Internet are explored.

Electronic mail (e-mail) is one of the most widely used applications of the Internet. Speed, ease of use, and no cost have made e-mail the communication of choice for millions of individuals around the world. Most browser software offers e-mail capabilities at no additional cost.

COMPUTER APPLICATIONS: AN OVERVIEW

If automobiles had developed as computers have, in the year 2004, a user would be able to buy a Mercedes Benz for under than $2, get over 2 million miles to a gallon, and park up to three dozen cars in the corner of an office! If airplanes had developed as computers have, today a user would be able to go around the globe in under 20 minutes for just 50 cents! The computer that weighed over 18 tons 50 years ago now weighs under 5 pounds. It is 100 times more powerful, and its cost is less than 1% the cost of the first computer (Bidgoli, 1993).

Every day most workers use computers directly or indirectly. For example, in schools, computers prepare transcripts and grade multiple-choice and true/false test questions. In grocery stores, computers manage the inventory systems and cash registers. Point-of-sale (POS) systems have significantly improved the efficiency and effectiveness of grocery chains and department stores. Scanners rapidly scan the universal product code (UPC) of different products and quickly come up with the price of the item being purchased, at the same time keeping track of the most current inventory status. Home appliances such as TVs, VCRs, microwaves, and cameras all use some type of computer.

A cashless and checkless society is just around the corner. Even now, a user can accomplish all his/her financial transactions by using computers. Computers assist bank customers in withdrawing money from accounts, depositing money to accounts, and transferring money from one account to another account, all transactions that could be conducted remotely by using a computer and the Internet.

Executives don't need to leave their offices to attend meetings in different cities, states, or countries. Through computer and video conferencing, an executive can attend meetings in different locations without ever leaving the office. In addition, reports can be prepared and transferred from location to location by means of computers and telecommunication systems. The space program could not have been achieved without computer help. Computers also have played a significant role in medical research and the treatment of various diseases.

Computers have significantly reduced the production time of textbooks and other print media. Authors routinely prepare various drafts of manuscripts on a computer with word processing software. Editors make changes based on comments from reviewers by using the editing features of the same software. Computerized typesetting and printing equipment then produce the book that a user reads. Warehousing, inventory control, shipping, and marketing are all accomplished with the help of computers. For example, the names and addresses of authors of this encyclopedia were stored in a computer file, and the mailing labels and letters sent to these individuals were prepared by means of a computer. Electronic mail kept the editor-in-chief of the encyclopedia in close contact with all the authors, scattered around the world. Several drafts of different chapters were sent back and forth between authors, editors, and reviewers all by using electronic mail. The sales and marketing of the encyclopedia have been significantly improved by using various applications of computers.

Nowadays, publishers can store the entire text of a book in several computer files. When new editions are needed, changes can be made easily by editing existing computer files instead of creating new ones.

Practically speaking, the number of applications performed by computers is unlimited. There is not a single day in which a worker is not affected by the computer. Nearly all business organizations have computers. All educational disciplines, even the humanities, have applications for computers. Every job now involves computers either directly or indirectly. Table 1 summarizes some of the more practical applications of computer technology.

DEFINING A COMPUTER

A computer is a machine that accepts data as input, processes the data without human interference using a set of stored instructions, and outputs information. Instructions are step-by-step directions given to a computer for performing specific tasks. Computer programs consist of instructions written in a language understood by a computer. There are hundreds of computer programs designed for different purposes. Several of them that are related to the Internet, and e-commerce applications are introduced in this encyclopedia.

Any computer system consists of hardware and software. Hardware components are all physical devices, such as a keyboard, monitor, mouse, data tablet, disk drive, zip drive, DVD (digital versatile disc) drive, CD-ROM, printer, modem, and processor unit. The software component consists of computer programs written in different computer languages. The software could be developed in-house or chosen from among the many thousands of commercial software packages readily available on the market for a moderate cost.

To communicate with a computer, a user must talk to it in its own language. There are hundreds of computer languages available. BASIC, FORTRAN (formula translator), COBOL (common business-oriented language), Java, and C++ are some example of computer languages.

A language used by many microcomputers is called BASIC (beginner's all-purpose symbolic instruction code). A simple example of a BASIC program follows:

```
10 A = 10
20 B = 20
30 C = A + B
40 PRINT C
50 END
```

Table 1 Popular Applications of Computers

Industry	Examples of Applications
Space program	Launching of spaceships
Medicine	Drug testing, monitoring vital signs, recording patient history
Geology	Earthquake analysis, soil and mineral analysis
Government	Tracking of Internal Revenue Service and social security information
Business	Accounting and finance for private and public organizations
Education	Computer-aided instruction, remote education sites, distance learning
Manufacturing	Computer-aided design, computer-aided manufacturing, robotics
Traffic control	Provision of more effective traffic systems
Appliances	Microwaves, VCRs, TVs, cameras
Music	Control and creation of sound
Entertainment	Generation of three-dimensional pictures and animation in movies and video
Airlines	Reservation systems, air traffic control systems, and navigation systems

This program instructs the computer to add values A and B, store the result in an address called C, and then print the result. Line 50 tells the computer that this is the end of the program.

The diagram presented in Figure 1 illustrates the building blocks of a computer. The input devices send data and information to the computer. The keyboard is one example of an input device. Other input devices may include mouse, touch screen, light pen, trackball, data tablet, bar code reader, optical character reader (OCR), and magnetic ink character reader (MICR). The output devices receive the output generated by the computer. The output may be presented on a display monitor, CRT (cathode ray tube), or VDT (video display terminal). This type of monitor output is called soft copy. If a printer generates the output, it is called hard copy. In some cases, plotters are used for generating hard-copy output as graphics.

The main memory (primary storage) is like the human brain. Memory is the location in which computers store data and instructions. Floppy disk, magnetic disk, magnetic tape, hard disk, CD-ROM, and zip disk are used as secondary storage devices. The CPU (central processing unit) is the heart of the computer. The CPU is divided into two components: the ALU and the control unit. The ALU (arithmetic/logic unit) performs arithmetic operations (addition, subtraction, multiplication, and division) and logical operations (e.g., comparing numbers). The control unit acts like the captain of a ship. It tells the computer what to do. For example, it tells the computer from which device the data is read or to which device to send the output. Figure 2 shows the components of a data processing system. Other types of memory include RAM, ROM, PROM, and EPROM.

Random access memory (RAM) is a volatile memory. Data stored in RAM will be lost in the event of a power failure. To avoid this type of loss, always, the user must save his/her work on a permanent memory medium (i.e., secondary memory), such as a diskette.

Read-only memory (ROM) is a prefabricated ROM chip supplied by vendors. This memory stores some general-purpose instructions or programs.

Programmable read-only memory (PROM) is memory a user can program by using a special device. However, once programmed, this type of memory cannot be erased.

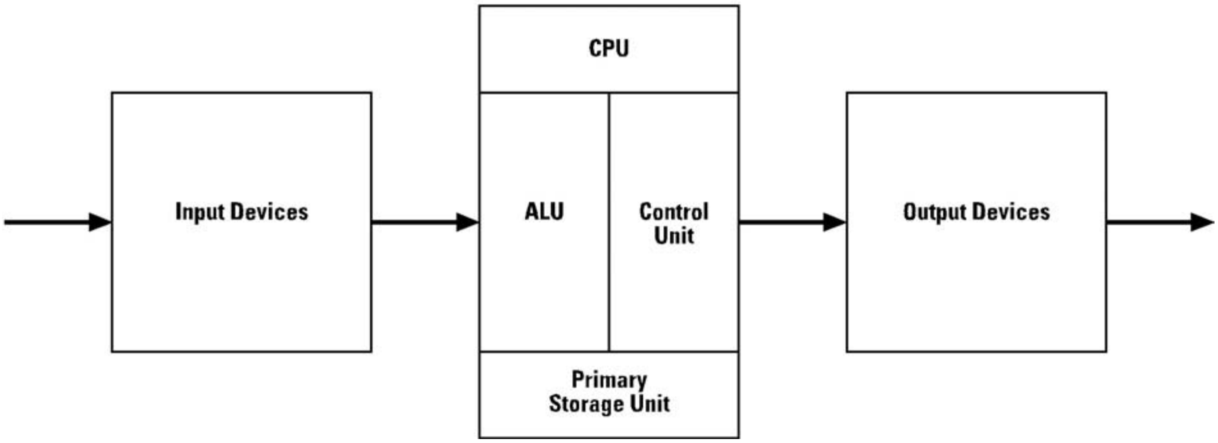


Figure 1: Building blocks of a computer.

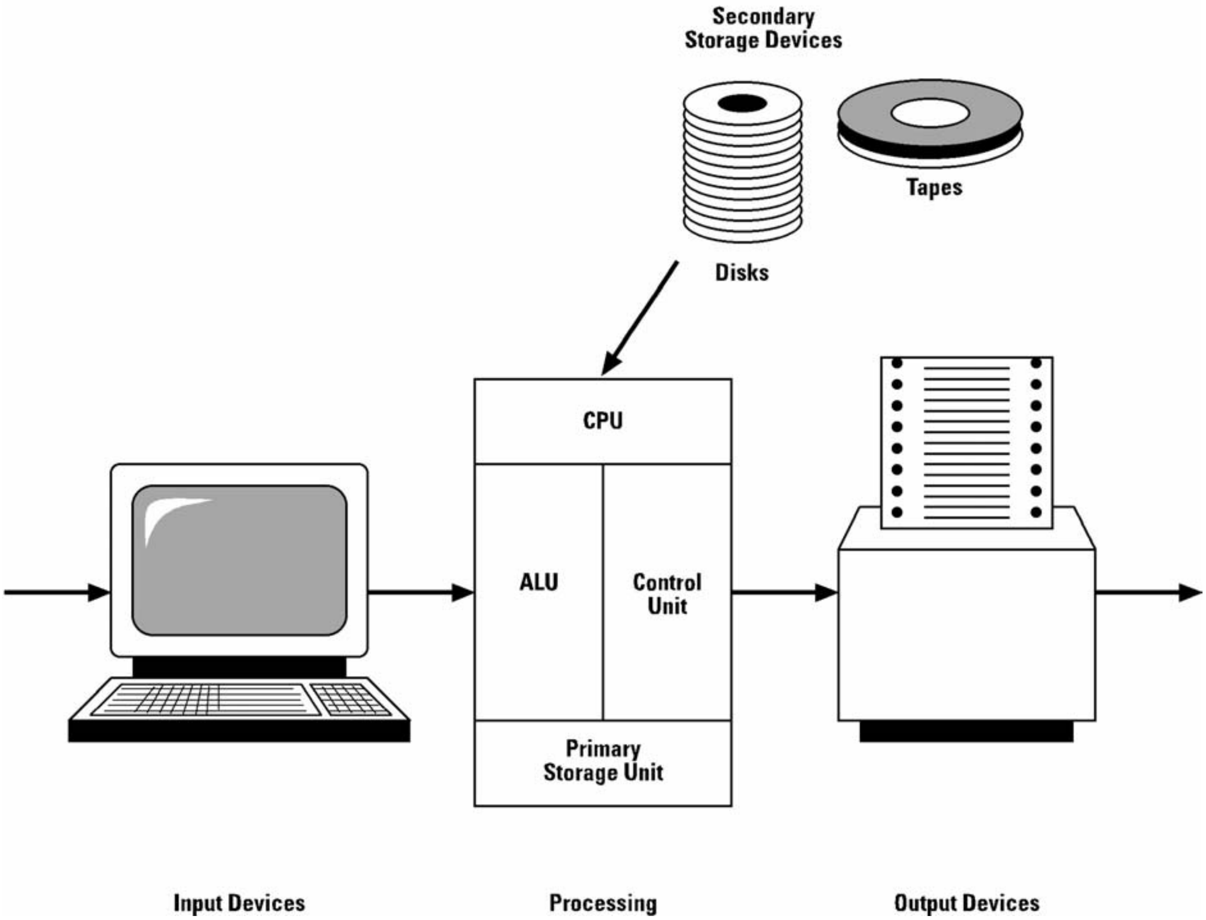


Figure 2: Components of a data processing system.

Erasable programmable read-only memory (EPROM) is read-only memory. It can be programmed by the user and, as the name indicates, erased and programmed again.

CATEGORIES OF COMPUTERS ACCORDING TO THEIR POWER

Computers can be classified in several ways. Usually computers are classified based on cost, memory size, speed, and sophistication (single tasking versus multitasking). A single-tasking computer performs one task at a time; a multitasking computer performs several tasks at a time. Based on these criteria, computers are classified into micro-, mini-, mainframe, and supercomputers. This means that supercomputers are more expensive and faster and possess much more memory than microcomputers, minicomputers, and mainframes. This justification applies to other classes of computers as well. Applications for these computers include everything from homework and home finances (microcomputer) to space shuttle launches (supercomputer). Because both the speed and the sophistication of microcomputers are steadily increasing, it is difficult to draw a clear line between mini- and microcomputer. The microcomputer of today has more power than the mainframe of the 1970s, and all indications suggest that this trend will continue.

CLASSES OF DATA PROCESSING SYSTEMS

Computers as we currently know them have been around for approximately 50 years. What did we do before computers existed? What do developing nations that do not have access to computers do? In these situations, data is processed with devices other than computers. The major classes of data processing systems include manual, mechanical, electromechanical, and electronic (see Figure 3).

Manual Data Processing

A manual data processing system processes data manually, meaning by hand. A user might use paper, pencil, and ledgers. Other devices include the Chinese abacus and the slide rule. Slow processing speed and low degree of accuracy are two major problems associated with this type of data processing.

Mechanical Data Processing

To improve on the inherent drawbacks of manual data processing, mechanical data processing was developed. The credit card imprinter is a mechanical data processing device. Imagine what would happened if every time

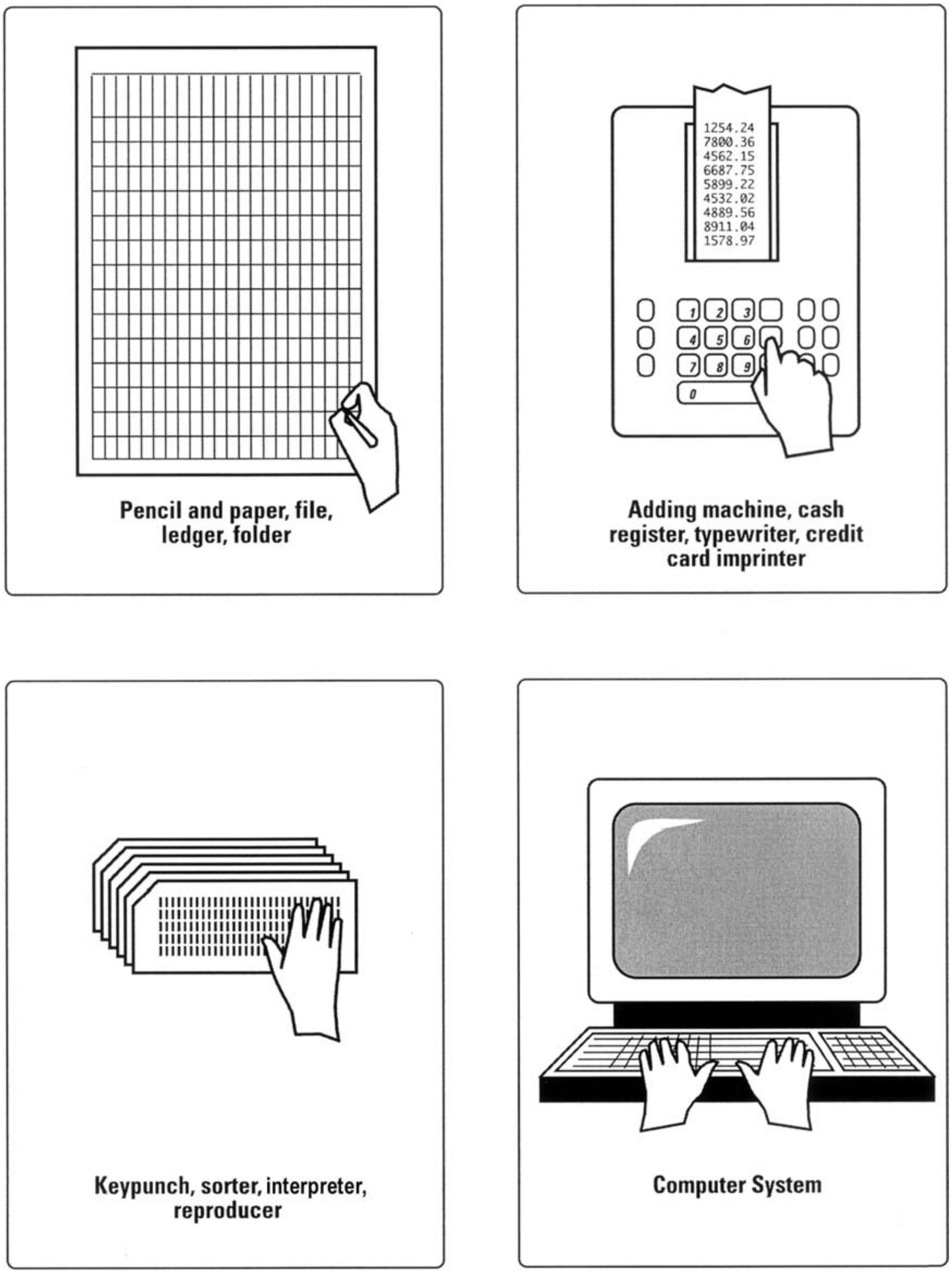

1254.24
7800.36
4562.15
6687.75
5899.22
4532.02
4889.56
8911.04
1578.97
Pencil and paper, file, ledger, folder
Adding machine, cash register, typewriter, credit card imprinter
Keypunch, sorter, interpreter, reproducer
Computer System

Figure 3: Data processing systems.

a user used his/her credit card to purchase gas, the operator in the gas station was required to write down all the information displayed on the card. Purchasing gas would be a tedious process, and the accuracy of the information could not be guaranteed. This is, in fact, still the case in some places. In more remote areas of the world, and in stores that do not process many credit cards, at times clerks manually write down credit card information.

Electromechanical Data Processing

Electromechanical systems differ from earlier systems of data processing in that the data or information must be

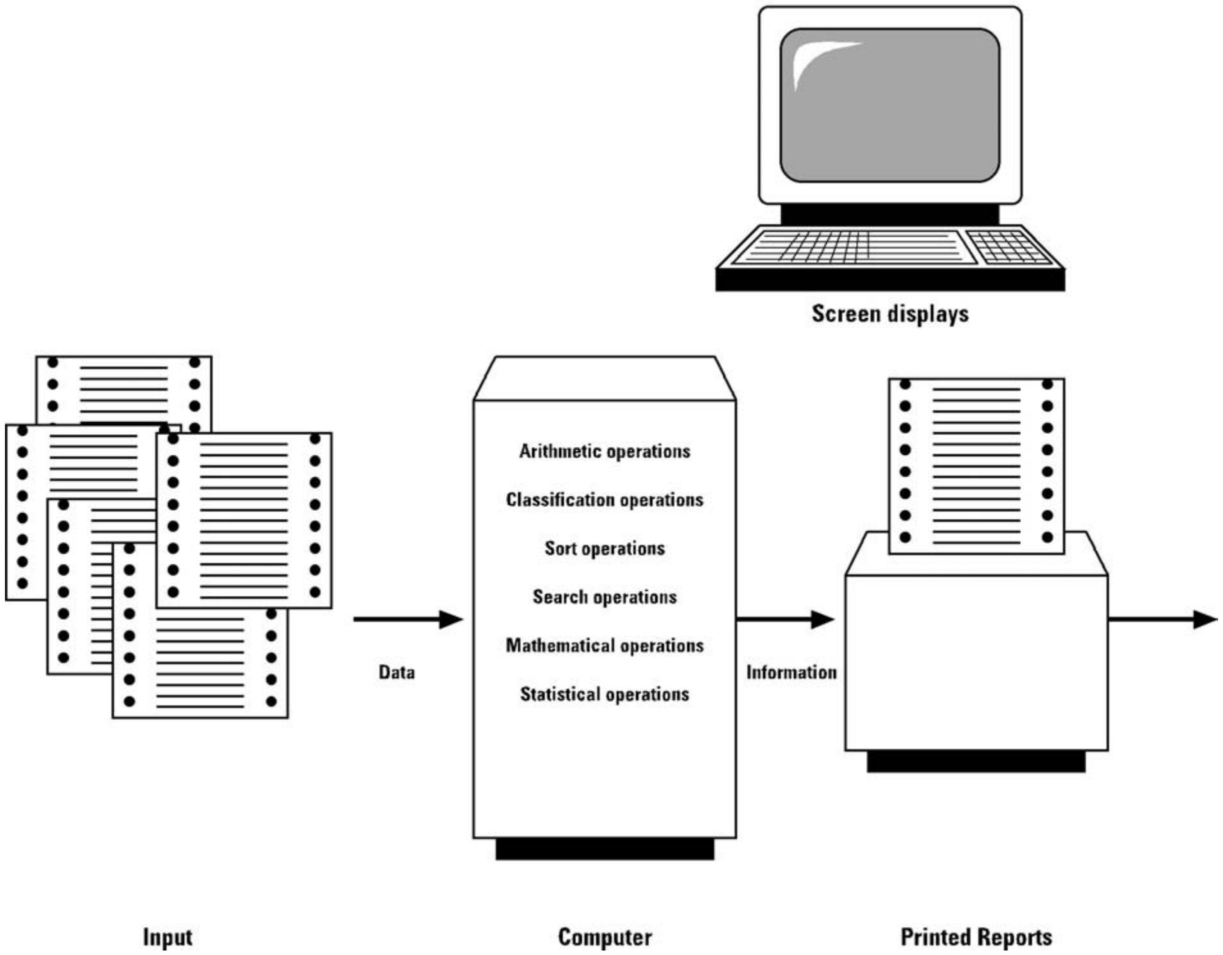


Figure 4: Data processing functions.

in machine-readable form. Punch cards, card readers, and sorters were invented to improve the speed and accuracy of data processing systems. By punching holes on a paper card, a punch-card machine puts data into a computer-readable format, which can be read very quickly. A sorter machine can sort thousands of cards with a high degree of speed and accuracy. A good example is the Hollerith's machine, which helped speed up the U.S. Census process in 1890.

Electronic Data Processing

Electronic data processing (EDP) is the fastest and most accurate of the data processing systems. In this type of data processing, data are entered to the computer through various devices, such as a keyboard or data tablet, and all processing is done electronically, without human intervention. As an example, consider a payroll processing system. The number of work hours and the pay rate of a worker are entered into the computer, his/her deductions are identified, and the computer prints a paycheck. The interesting aspect of EDP is that the same computer program that prints a paycheck for one employee is able to print paychecks for thousands of employees as soon as their employment data are entered. Throughout this encyclopedia, many examples of the amazing power, sophistication, and diverse applications of electronic data processing systems are presented.

DATA VERSUS INFORMATION

Figure 4 shows a simple diagram for a data processing system. The input to a data processing system is raw data. In the payroll example introduced earlier, number of hours and pay rate are examples of raw data. Looking at these two numbers individually, a user will not know how much this employee will earn in a given period, where he/she stands compared with the rest of the employees as far as salary is concerned, and so forth. A computer processes the data, and the results are called information or processed facts. In the payroll example, a paycheck may indicate an example of information.

What is the difference between data and information? Data is raw facts. If a user has only facts (data), he/she cannot make a decision. As an example, consider $210 million sales generated by Company X in their last fiscal quarter. Is this data or information? Using the same analogy as the payroll example above, we could conclude that, again, this is data, because a user will not be able to say anything regarding the performance of the company. Is this a profitable company? Did the company meet their sales objective? Did the sale increase or decrease?

Just looking at this number could raise many other unanswered questions. The entire field of information systems is concerned with the production of timely, accurate, and useful information. Many organizations are data rich but information poor. The real challenge for practitioners in the dynamic field of information processing is to design computer-based information systems to generate accurate, timely, and useful information.

The information needed by a user directly affects the type of data used in an information system. If an organization has defined its strategic goals, objectives, and critical success factors to ensure a viable and growing organization, the data component can be structured rather easily and the information system has potential for success. On the other hand, if there are conflicting goals and objectives, or if the company is not aware of which factors are critical to its success, many problems can occur to destroy confidence in its information system or minimize the system's effectiveness. If the information system is not designed to evolve as changes (both internal and external) take place, then the system may do more damage than good.

Of course, the objectives of the organization ultimately resolve the questions of the sources of data—external or internal sources—and whether the data is past (performance), present (operational), or future (budget or cash flow) oriented. The urgency of need and the availability of data in many forms, including aggregated (lump sum) or disaggregated (itemized), can then be addressed. Disaggregated data is needed when, for instance, sales are analyzed by product, territory, or salesperson, and costs are analyzed by cost center or product. Aggregated data limits the decision maker's ability to focus on specific factors.

DEFINING DATA PROCESSING

When data has been processed to generate information, one or several of the following tasks may have taken place.

Arithmetic Operations

A user might add all the sales generated by each region in the northwest to generate a value that indicates the total sales of a company in the northwest region.

Sort Operations

A user might sort the total sales of a company for all 50 states from the highest to the lowest volume to find the best and the worst sales regions.

Classification Operations

A user might classify a sales data into 10 different product groups. Using this information, a user can see the total sales generated by product group X. This information will help a user choose the best and the worst products in a company on the basis of total sales.

Search Operations

A user may search a database in order to find a specific item that meets certain criteria, for example searching for a sales region that has generated record sales in three or more periods.

Statistical and/or Mathematical Operations

A user may utilize statistical and/or mathematical models in order to generate a sales forecast for the next period. This forecast may serve as the basis for sales planning for the next period.

Figure 4 illustrates data processing functions.

There are many types of data processing that enable a user to convert raw data to information. The method a user chooses depends on the specific situation. Remember, however, that the computer processes only the data it is given. If the data is erroneous, the information provided will be erroneous as well. A user should always remember the principle "garbage in, garbage out" (GIGO). Experts in the field call computers error-free machines, given establishment of proper working conditions. Proper working conditions include factors such as proper room temperature (not too hot or too cold), a dust-free environment, appropriate humidity, and a virus-free computer.

THE POWER OF COMPUTERS

Computers draw their power from three distinguishing factors that far exceed the capacity of any human being: tremendous speed, high degree of accuracy, and immense storage and retrieval capabilities.

Speed

Computers process data with amazing speed. Speed is measured as the number of instructions performed per second, as follows:

Millisecond	=1/1,000 second	=1 thousandth
Microsecond	=1/1,000,000 second	=1 millionth
Nanosecond	=1/1,000,000,000 second	=1 billionth
Picosecond	=1/1,000,000,000,000 second	=1 trillionth

Early microcomputers (1981) operated at a speed of 4.7 megahertz (MHz or sometimes Mhz, a million cycles per second). This is how a microprocessor clock speed is measured. In the year 2003, microcomputers with the processing speeds of 2.4 MHz are readily available.

Accuracy

Computers are error-free machines—they do not make mistakes. To make the accuracy issue more clear, consider the following two numbers:

2.00000000001
2.00000000002

To humans, these numbers are so close that they could be considered equal. To a computer, however, these two numbers are very different. For some computer applications, such as those used in space missions, this level of accuracy and speed is required to bring the space shuttle back to Earth at a given time and in a specific location.

Storage and Retrieval

Storage means saving data or information in the memory of a computer. Retrieval is the act of bringing the data or information back from memory. Computers can store vast quantities of data and can locate specific items very quickly.

COMPUTER OPERATIONS

Computers can perform three basic tasks: arithmetic operations, logical operations, and storage and retrieval operations (as explained above). One or a combination of these tasks accomplishes all other tasks. For example, playing games could be a combination of all three functions.

Arithmetic Operations

Computers can add, subtract, multiply, divide, and raise to power. The five basic operations are as follows:

A + B (addition)	5 + 2 = 7
A − B (subtraction)	5 − 2 = 3
A * B (multiplication)	5 * 2 = 10
A/B (division)	5/2 = 2.5
AˆB (exponentiation)	5ˆ2 = 25

Logical Operations

Computers can perform logical operations by comparing two numbers. For example, a computer can compare *A* and *B* to determine which number is larger.

COMPUTER GENERATIONS

Over the past 5 decades, there have been major advancements in hardware. Computers began with vacuum tube technology. In 1946, rudimentary computers were bulky and unreliable. They generated excessive heat and were very difficult to program. The second generation of computers began in 1957. This generation, which used transistors, was indeed a significant improvement over the first. These computers were faster, more reliable, and easier to program and maintain. The third generation began in 1964 with computers that operated on integrated circuits (IC), which enabled computers to be even smaller, faster, more reliable, and more sophisticated. Remote data entry and telecommunications were introduced during this generation. The fourth generation began about 1970. This generation of computers is associated with several attributes: miniaturization, very-large-scale integration (VLSI) circuits, ultra-large-scale integration (ULSI), and widespread applications of microcomputers, optical disks, and bubble memories. Bubble memory is built on a thin crystalline film (mineral garment). Through polarization of the bubbles, data is presented on this nonvolatile memory. The presence of a bubble represents a one, and the absence of a bubble represents a zero. Two drawbacks of this kind of memory are its high cost and its relatively low speed.

The early 1990s heralded the beginning of the so-called fifth generation of computer technology. The major attributes of this generation include parallel processing, gallium arsenide chips, and optical technologies. A parallel processing computer contains hundreds or thousands of CPUs, which means that the computer is capable of processing data much faster than its predecessors.

Because silicon technology is not able to emit light and has speed limitations, computer designers have concentrated on gallium arsenide technology. Electrons move almost five times faster in gallium arsenide than they do in silicon. Devices made with this synthetic compound can emit light and withstand higher temperatures and survive much higher doses of radiation than silicon devices.

The major problems associated with gallium arsenide are the difficulties in mass production and working with it. Gallium arsenide is soft and fragile compared with silicon; it breaks more easily during slicing and polishing. Currently, because of high costs and difficulty of production, military systems are the major users of this technology. However, research continues to eliminate some of the shortcomings of this impressive technology.

Optical technologies offer at least three features not found in earlier technologies: greater speed, parallelism (several thousand independent light beams can pass through an ordinary device), and interconnection (denser arrays of interconnections are possible because light rays do not affect each other). Optical computing is in its infancy and much more research is needed to produce a full-featured optical computer. Nevertheless, the storage devices using this technology are revolutionizing the computer field by enabling massive amounts of data to be stored in very small spaces.

The fifth generation will include revolutionary architecture that did not exist in the first four generations of computer technology. Fifth-generation computers work with so-called artificial intelligence (AI) technologies. AI is expected to make the computer smarter—so much so that AI computers are expected to do some of the tasks that humans perform. Several AI products are now on the market. Among them are experts systems, robotics, and natural language processing. Table 2 highlights the trends in hardware technology.

In parallel with hardware technology, software technology has also improved significantly. When the first computers were introduced, in the early 1940s, the only language understood by these computers was machine language. Machine language consists of a series of zeros and ones. It is difficult to program a computer using machine language. At that time, the only users of computers were highly trained data processing personnel. The second generation of computer language was called assembly language, and it consisted of a series of short codes, or mnemonics. One can say, in general, that

Table 2 Hardware Trends

Generation	Date	Major Attribute
First	1946–1956	Vacuum tubes
Second	1957–1963	Transistors
Third	1964–1970	Integrated circuits
Fourth	1971–1992	VLSI
Fifth	1992–?	Gallium arsenide, parallel processing

Table 3 Software Trends

Generation	Attribute
First	Machine language
Second	Assembly language
Third	High-level language
Fourth	4GL
Fifth	NLP

assembly language is easier to use and understand than machine language. But, still, rigorous training is needed to become the proficient user of a computer using this language. The third generation of computer languages, called high-level languages, or higher level languages, is more user- and application-oriented. A language like COBOL is self-explanatory.

A user can get a general idea when he/she reads a program written in COBOL. These languages are easier than the first two to learn and use. However, they are still very specific in nature and each is more suitable for one type of application than another. The fourth-level computer languages, or 4GLs, are a lot more forgiving than the first three. They are by far the easiest to learn and use. Much less time is needed to perform a task using these languages compared with their earlier counterparts. However, to use even these languages, the user has to have some basic computer training and should be familiar with the keyboard. The fifth generation of computer languages, natural language processing (NLP), promises a great deal of flexibility and power. If these languages ever become a reality, computers will become a lot easier to use, friendlier, and more flexible. Table 3 highlights the software trends.

INPUT, PROCESS, AND OUTPUT CONCEPTS

Most users and decision makers use a computer as a black box. This means that a user sends data to the computer, the computer performs the processing task, and information is given back to the user on some type of output device. According to the black box theory, the user doesn't need to know what goes on inside the "box." The user provides the input data, tells the computer how this data should be processed, and leaves the rest to the computer. However, to be able to effectively use a computer, basic knowledge of input, processing, and output is essential.

DATA REPRESENTATION

Every character, number, or special symbol that a user types on the keyboard is represented as a binary number in the computer's memory. A binary system consists of 0 and 1. The 1 represents "on" (high voltage), and the 0 represents "off" (low voltage). Figure 5 illustrates this system. To represent these different data items, computers use special formats.

Computers and communications systems use data codes in order to represent and transfer data between various computers and network systems. Three popular types of data codes are (Bidgoli, 2000) Baudot code, ASCII (American standard code for information interchange), and EBCDIC (extended binary coded decimal interchange code).

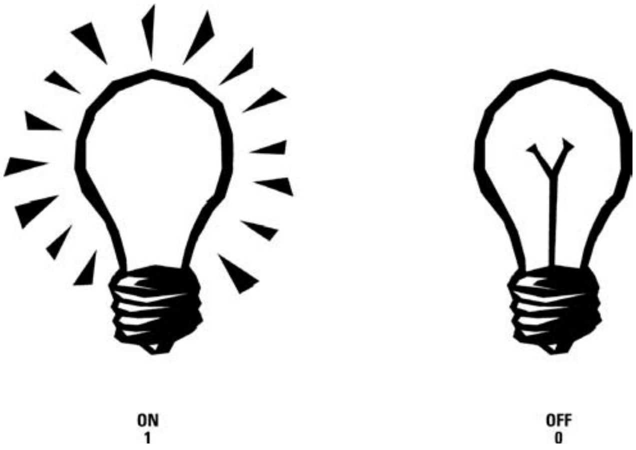


Figure 5: Graphic representation of a binary system.

The *Baudot code* was named after a French engineer, Jean-Maurice Emile Baudot. It was first used to measure the speed of telegraph transmissions. It uses 5-bit patterns to represent the characters A to Z, the numbers 0 to 9, and several special characters. Using Baudot code up to 32 characters can be defined ($2^5 = 32$). This is not enough to represent all the letters of alphabet (uppercase and lowercase) and special characters. To overcome this limitation, Baudot code uses DOWNSHIFT (11111) and UPSHIFT (11011) character code. By doing this, up to 64 different characters can be defined, doubling the code's original size. This process is similar to typing using a keyboard. When a user presses the Caps Lock key on the keyboard, turning on the Caps Lock feature, all the characters typed are transmitted as uppercase. As soon as the user presses the Caps Lock key again, undoing the feature, all the characters are transmitted as lowercase.

Therefore, using Baudot, every character except for a space is either an upshift character or a downshift character. For example, 10001 represents both the letter Z and the + (plus sign). The letter Z is a DOWNSHIFT character and the plus sign is an UPSHIFT character. Baudot code is no longer used. However, it illustrates how information is transmitted by a small number of bit combinations.

The American National Standards Institute (ANSI) developed ASCII (pronounced "ask-ee"). It is the most common format for text files, for PC applications, and on the Internet. In an ASCII file, each alphabetic, numeric, or special character is represented with a 7-bit binary number (a string of seven zeros or ones). Up to 128 characters can be defined ($2^7 = 128$). UNIX- and DOS-based operating systems and a majority of the PC applications use ASCII for text files. Windows NT (2000) uses a newer code, called Unicode. The extended ASCII is an 8-bit code used by IBM mainframe computers, which allows 256 (2^8) characters to be represented.

EBCDIC (pronounced "ehb-suh-dik") is a binary code for alphabetic and numeric characters that IBM

Table 4 Selected Keyboard Symbols in ASCII and EBCDIC

Symbol	ASCII	EBCDIC
Space	0100000	01000000
A	1000001	11000001
B	1000010	11000010
Z	1011010	11101001
a	1100001	10000001
b	1100010	10000010
z	1111101	10101001
*	0101010	01011100
%	0100101	01101100
(	0101000	01001101
0	0110000	11110000
1	0110001	11110001
9	0111001	11111001

developed for its mainframe operating systems. It is the code for text files that is used in IBM's OS/390 operating system running on its S/390 servers, used by many corporations. In an EBCDIC format, each alphabetic, numeric, or special character is represented with an 8-bit binary number (a string of eight zeros or ones). Up to 256 characters (letters of the alphabet, numerals, and special characters) can be defined ($2^8 = 256$). Conversion programs allow different operating systems to convert a file from one code to another. Table 4 illustrates selected examples of ASCII and EBCDIC.

TYPES OF PROCESSING

Computer users can process data either in batch mode or in real-time mode. In batch processing, data is sent to the computer periodically, for example every 24 hours, every 2 weeks, or every month. This type of processing is suitable for applications that don't need an immediate response. Payroll is a good example—hourly employees do not need to be paid each hour.

Many applications, however, require an immediate response. In such applications, a transaction is processed as soon as it occurs. This type of processing is known as real-time processing, or transaction processing. An airline reservation system is a good example of this type of processing. A reservation must be entered into the computer immediately, otherwise one seat may be sold to many customers or a flight may have many empty seats.

DIFFERENT CLASSES OF COMPUTER SOFTWARE

Figure 6 illustrates important software used in a computing environment.

In broad terms, software is divided into two major groups: application software and system software. Application software is associated with what users routinely work with on a daily basis to get their work accomplished. Examples of application software include word processing, spreadsheet, database, desktop publishing, financial planning, and project management. System software is commonly associated with the efficient and reliable functioning of the computer system. Examples of system software include operating system software, utility, and compiler. An operating system is a set of programs that controls and supervises computer hardware and software. Initially, the utility software market niche developed in order to "fill the gaps"—the missing or desired capabilities of operating systems. Since that time, an entire software industry has developed to meet the demand. Some of the utility software applications were developed to give users capabilities that were not available in operating systems (Bidgoli & Prestage, 2003). Other utility software applications were developed to make using operating systems simpler. Virus protection software can be regarded as utility software in that although it does not directly serve the user's information processing needs, it protects the operating system, applications, and data from harm caused by malicious virus applications. A compiler is a program that translates a user program from source code into object code.

To a typical user, application software is more important and more commonly used than other types of software. A microcomputer can perform a variety of tasks by using either commercial software or software developed in-house. Software developed in-house is usually more expensive than commercial software. However, such software is more customized and may better fit the users' needs. There are several thousands of software packages available for PCs. For almost any task a user can think of, there is an appropriate software package on the market. The following are among the most popular productivity packages and applications available for microcomputers.

Word Processing

A microcomputer used as a word processor is similar to a typewriter with a memory. With such a facility, a user can generate documents, reports, and brochures, make deletions and insertions, and cut and paste. Word processing programs are becoming more sophisticated. Some of these programs provide graphics and data management features. With word processing programs, hundreds of hours can be saved by not having to type the same document repeatedly. Organizations often send the same letter to many of their customers. The only differences in these letters are the names and addresses of the customers. Word processing programs include a mail merge feature that allows and expedites mass mailing.

The majority of word processors now include spell checkers that are able to correct most of the misspelled words in a document. The next challenge is the creation of documents that include correct verbs, subjects, and adjectives and a smooth style. Grammar checkers are able to correct a document for grammatical errors. Also, the creation of simple, easy-to-read sentences is of prime importance. Grammar-checking software promotes good writing.

Grammar checkers perform text analyses by using linguistic analysis, parsing, and rule matching. A parser parses, or simply put, it breaks long sentences into shorter ones. Grammar checkers play an especially important role when multiple authors are involved in a project. In such

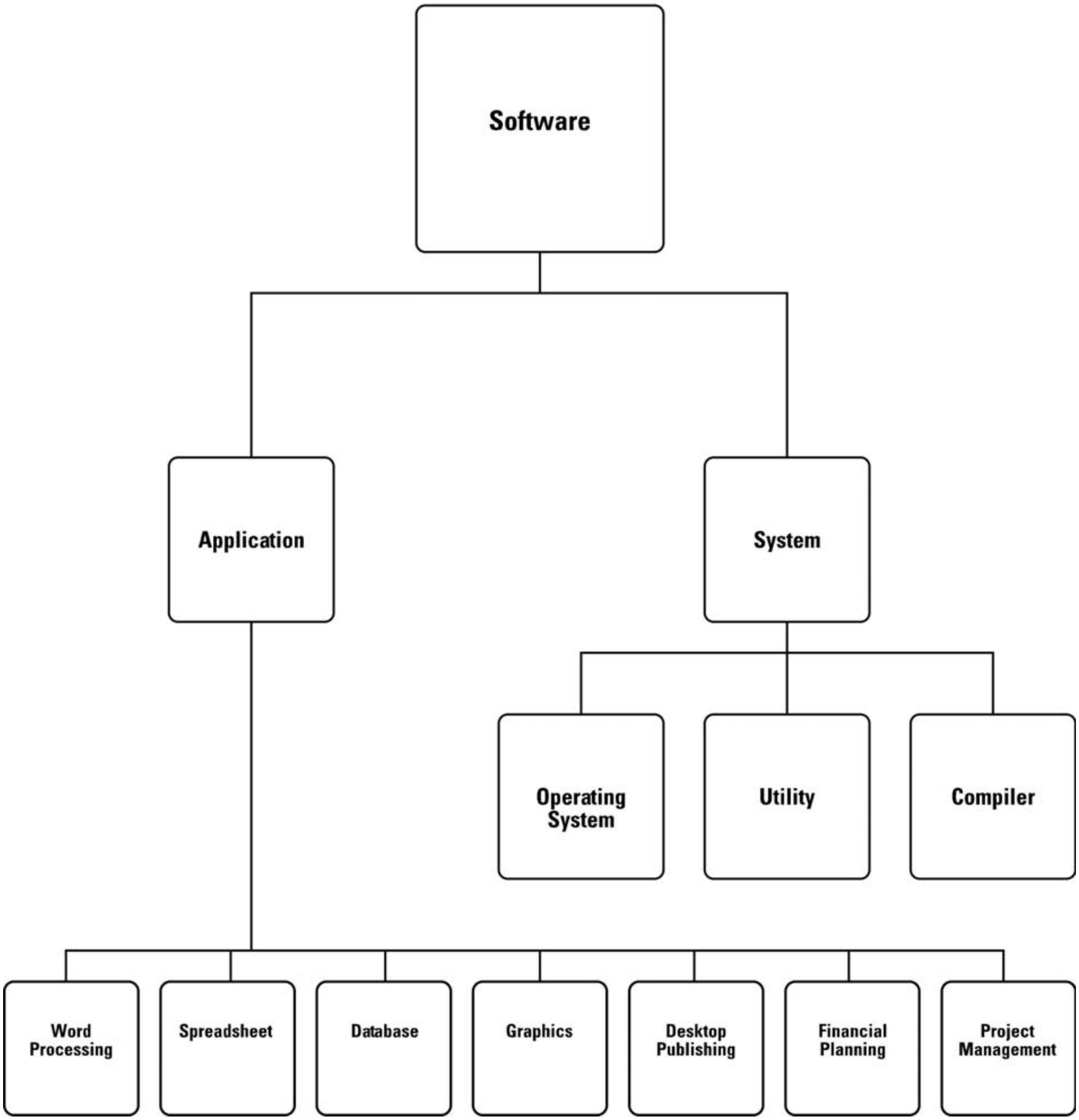


Figure 6: Important software used in a computing environment.

cases, grammar-checking software creates uniformity of tone, reading level, and style. Grammar-checking software is not 100% perfect yet, but it has come a long way.

Spreadsheet

A spreadsheet is a table of rows and columns. Typical spreadsheet software is capable of performing numerous tasks. Microsoft Excel, for example, is capable of performing spreadsheet functions as well as database and graphics functions. The number of tasks that can be performed by a spreadsheet program is almost unlimited. Any application suitable for row and column analysis is a candidate for a typical spreadsheet. For example, a user may use a spreadsheet to prepare a budget. When he/she is done with the budget, some impressive "what-if" analyses can be performed. This means a user may manipulate variables on the spreadsheet, for example reducing the income by 2% and directing the spreadsheet to calculate the effect of this change on other items on the spreadsheet. Or a user may want to see the effect of 2% reduction in the interest rate on a house mortgage.

Database

Database software is designed to perform database operations, such as file creation, deletion, modification, search, sort, merge, and join (combining two files or tables based on a common key). A file is a collection of records. A record is a collection of fields. A field is a collection of characters.

A database can also be compared to a table of rows and columns. The rows correspond to records, and columns correspond to the fields (attributes) within the record. As an example, the ABC Company keeps track of its customers in a database. This database, or table, has 5,000 rows (records) and five columns (fields). The columns keep tracks of names, credit limit, address, income, and years of service. Common applications of database software are sorting and searching records. In sort operations, the user enters a series of records in any order, then asks the database management program to sort the records in ascending or descending order based on the data in the fields. Search operations are more interesting. A user can search for data items that meet certain criteria, for example all customers who have an income greater than $80,000, who live in the southwest, and who have been customers for more than 5 years. Some database management systems allow a user to search for key words within a text file. This type of application is common when a user performs literature research.

Graphics

Graphics software has been designed to present data in graphic format. Data can be converted into a line graph to show a trend, into a pie chart to highlight the components of a data item, and into other types of graphs for various analyses. Masses of data can be converted to a graph, and in an instant, the user can discover the general pattern of the data. Different graphs achieve different tasks. For example, a pie chart shows the components of a data item. It may break down the total cost into raw material costs, labor costs, and advertising costs, as an example. A line graph may show the sales trend of the past 5 years and yet a bar graph compares the profit of the ABC Company for the past 6 years. Graphs can easily highlight patterns and the correlation among data items. They also make data presentation a more manageable task. Graphics are produced either by integrated packages, such as Microsoft Excel, or by dedicated graphics packages, such as Harvard Graphics, by the Harvard Graphics Corporation.

Desktop Publishing

Desktop publishing is used to produce professional-quality documents (with or without graphics) using relatively inexpensive hardware and software. All that is needed are a PC, desktop publishing software, and a laser or letter-quality printer. Desktop publishing has evolved as a result of three major factors: (a) inexpensive PCs, (b) inexpensive laser printers, and (c) sophisticated and easy-to-use desktop publishing software.

Desktop publishing enables a user to produce high-quality screen output and then transfer it to the printer, using the “what you see is what you get,” or WYSIWYG, concept. Major applications of desktop publishing are for creating newsletters, brochures, training manuals, transparencies, posters, and books.

Financial Planning

Financial planning software works with large amounts of data and performs diverse financial analyses. These analyses include present value, future value, rate of return, cash flow analyses, depreciation analyses, retirement planning, and budgeting analyses. Using financial planning software, a user can plan and analyze a financial situation. For example, a user will know how much a $2,000 IRA will be worth at 6% interest in 30 years, and he/she can translate the value of future cash flows into today’s dollars. A user will know how much he/she has to deposit in a bank in order to have $150,000 in 18 years for a child’s education.

Project Management

A project consists of a series of related activities. Building a house, designing an order entry system, and writing theses are examples of projects. The goal of project management software is to help a project manager keep time and budget under control by resolving scheduling problems. Project management software helps managers plan and set achievable goals. It highlights the bottlenecks and the relationships between different activities. This software allows the user to study the cost, time, and resource impact of any change in the schedule.

TAKING CARE OF A MICROCOMPUTER

Maintaining a microcomputer in a proper working condition requires the following:

Protection against dirt, dust, and smoke.

Making backup copies of data for security reasons and keeping backups in different locations in case of data loses, computer crashes, fires and floods.

Avoiding any kind of liquid spills.

Maintaining steady power; using surge protectors for power fluctuations and using lightning arresters in mountainous areas.

Protecting the system from static by using humidifiers or antistatic spray devices.

Not starting a computer using an unfamiliar disk (avoiding computer viruses—dangerous programs that erase and/or corrupt all data).

Not downloading information to a computer from unknown bulletin boards (downloading means to import information from other computers and network systems to the user’s system using a modem and telephone line).

Not opening e-mail attachments that are not expected or that are sent from unknown senders.

Acquiring insurance for computer equipment.

Keeping track of the serial numbers of all computer equipment and software; also, keeping a phone list of all vendors and help desk phone lines and all maintenance contracts.

CONCLUSION

This chapter provided an overview of computer literacy. It explained different types of data processing and the differences between data and information. The unique characteristics of computers, different applications, and different generations of computers were introduced. The chapter reviewed various types of computer memories and basic productivity tools. It reviewed batch processing and transaction processing systems and concluded with guidelines for the proper maintenance of a microcomputer system.

GLOSSARY

Computer A machine that accepts data as input, processes the data without human interference using a set of stored instructions, and outputs information. Instructions are step-by-step directions given to a computer for performing specific tasks.

Computer Generations Different classes of computer technology identified by a distinct architecture and technology; the first generation was vacuum tubes, the second transistors, the third integrated circuits, the fourth very-large-scale integration, and the fifth gallium arsenide and parallel processing.

Computer Literacy The basic knowledge needed for understanding and using the computers for day-to-day tasks.

Erasable Programmable Read-Only Memory (EPROM) A ROM that can be programmed by the user and, as the name indicates, erased and programmed again.

Programmable Read-Only Memory (PROM) Memory that, by using a special device, the user can program. However, once programmed, this type of memory cannot be erased.

Random Access Memory (RAM) A volatile memory that stores data and instructions during a session.

Read-Only Memory (ROM) A prefabricated chip supplied by vendors. This memory stores some general-purpose instructions or programs.

CROSS REFERENCES

See *Electronic Commerce and Electronic Business; Internet Literacy; Internet Navigation (Basics, Services, and Portals).*

REFERENCES

Beekman, G. (2003). *Computer confluence* (5th ed). Upper Saddle River, NJ: Prentice Hall.

Bidgoli, H. (1993). *Information systems literacy: Concepts, DOS 5.0, WordPerfect 5.1, Lotus 1-2-3 (2.3), and dBase IV (1.1and 1.5)*. New York: MacMillan Publishing Company.

Bidgoli, H. (1999). *Handbook of management information systems: A managerial perspective*. San Diego, CA: Academic Press.

Bidgoli, H. (2000). *Handbook of business data communications: A managerial perspective*. San Diego, CA: Academic Press.

Bidgoli, H. (2002). *Electronic commerce: Principles and practice*, San Diego, CA: Academic Press.

Bidgoli, H., & Prestage, A. (2003). Operating systems. In H. Bidgoli (Ed.), *Encyclopedia of information systems* (Vol. 3, pp. 377–390). San Diego: Academic Press.

Haag, S., Cumming, M., and Rea, A. I., Jr. (2002). *Computing concepts*. Boston: McGraw–Hill Irwin.

Information literacy on the WWW. Retrieved February 1, 2003, from http://www.fiu.edu/~library/ili/iliweb.html

Online encyclopedia. Retrieved February 1, 2003, from http://whatis.com

Online encyclopedia. Retrieved October 4, 2002, February 1, 2003 from http://webopedia.com/

Parsons, J. (2002). *Computer concepts—Illustrated introductory* (4th ed.). Boston: Course Technology.

Computer Security Incident Response Teams (CSIRTs)

Raymond R. Panko, *University of Hawaii at Manoa*

INTRODUCTION

Almost all corporations today protect themselves with layered defenses consisting of firewalls, antivirus systems, hardened hosts, and other protections. Even so, security incidents (also called security breaches) sometimes occur.

The firm's on-duty staff may be tasked to handle minor incidents because they can respond immediately and generally effectively. For major incidents, however, such as a major virus attack, a major denial-of-service attack, or the hacking (take-over) of important servers, the firm needs a team approach to stop the breach and get the firm back to normal. To handle major incidents, many firms create computer security incident response teams (CSIRTS), also known as computer emergency response teams (CERTs).

A critical success factor for any CSIRT is rapid response. During a major security breach, a corporation's operations are likely to be disrupted. This can result in the loss of immediate revenues, in the imposition of penalties by business partners and regulators, and by a loss in customer or investor confidence. During the intense stresses of major security incidents, CSIRTs must be able to act very rapidly but correctly.

During emergencies, human beings often are not at their best cognitively. Instead of ensuring that they understand the situation well, they often fixate on a single possible solution that eventually turns out to be fatally flawed. They also tend to make mistakes and work slowly when they are handling complex and difficult tasks for the first time. In addition, teams whose members are unfamiliar with the way their coworkers operate often have communication breakdowns.

Consequently, it is extremely important for CSIRTs to be organized and to conduct live rehearsals before incidents occur. Although first attempts may not be satisfactory, live rehearsals will identify problems that can be corrected so that performance is good during real incidents.

Once a real security breach occurs, the CSIRT's work must proceed through a series of well-considered steps (Panko, 2004). These are the discovery of an incident and its escalation to the status of major incident, the analysis of the incident to understand it, the containment of the attack to stop the spread of damage, repair to get damaged systems cleaned up and operational, the possible punishment of the attacker, the hardening of damaged systems against new attacks, and postmortem analysis.

BEFORE THE INCIDENT

Justifying the CSIRT

CSIRTs are expensive. It is important to be able to justify the cost of having a CSIRT to senior management. It is also important to be able to justify requiring participation from a number of departments, including during live rehearsals, which are very time consuming.

The normal way to justify a CSIRT is to collect data on the frequency of attacks and the damage caused by attacks. One excellent general source is the Computer Emergency Response Team/Coordination Center (http://www.cert.org).

Other good general sources are SecurityFocus (http://www.securityfocus.com) and SANS (http://www.sans.org). The FBI and the Computer Security Institute conduct an annual survey of corporations to assess the frequencies of various kinds of attacks (http://www.gocsi.com). A good source of data on viruses and spam is MessageLabs (http://www.messagelabs.com). In presenting summary information to top management, it is good to have data on both general trends and some examples of serious problems caused by incidents in specific corporations.

This information should be combined with estimates of how much major attacks of various types would cost the firm and, if possible, how having rapid detection and containment through the use of a CSIRT could reduce these costs.

Organizing the CSIRT

The first step is to create the CSIRT. Information technology (IT) security incidents can affect several functional groups in a firm. It usually also takes the staff from several functional areas to handle major IT security incidents. Consequently, the CSIRT must draw on several functional areas for its membership (Panko, 2004): the IT security staff, the IT staff, functional line managers, public relations, the legal department, and external organizations, such as outsourcers and law enforcement agencies.

The IT Security Staff

Obviously, the IT security staff needs to be involved and may lead the CSIRT. Most firms have very small IT security staffs, and when this is the case, it is normal for the entire IT security staff to be on the CSIRT.

The IT Staff

Sometimes, the IT security staff is under the information technology director. This is not ideal because a large fraction of all IT security breaches are committed by the IT staff. It is better for accountability for IT security to be a parallel operation. Another reason for keeping the IT security staff separate is that the IT staff sometimes focuses primarily on technology, while the IT security staff must focus more broadly on organizational issues. On the negative side, separating IT security from IT can result in friction between the two groups in terms of the ownership of server logs and other important matters. In addition, a separated IT security staff cannot "order" the IT staff to take action, and separation frees the IT department from accountability for security.

In CSIRTs, it is important to have members of the IT staff be involved however IT security and IT relate organizationally. It is also important to have operational levels of IT represented as well as IT management because operational employees often have a better picture of what is happening at a detailed level, while IT managers have the broader perspective needed to understand what is happening across systems and needed to understand the corporate implications of IT security breaches and alternative remedies.

Functional Line Managers

Fixing an IT security breach is not simply a technical matter. For example, the simplest and most effective way to contain a hacking intrusion is to terminate a victim computer's network or Internet connection. However, the business implications of "pulling the plug" can be horrendous. It is important for functional line managers in critical areas to be involved in and to approve the CSIRT's actions.

Public Relations

One non-IT group that needs to be strongly involved is the company's public relations department. If there is a serious problem, the company will want to manage communication with the outside world to produce as little damage as possible to public relations. Under no circumstances should an IT or IT security employee talk to the media directly. Many firms also work with external public relations companies, and there are even public relations companies that specialize in incident handling.

Legal Department

It is critical to have members of the legal department on the CSIRT. Lawyers have the specialized knowledge needed for prosecution and to defend the company against liability. (If an attacker compromises a corporate computer and uses it to attack another corporation, the company whose computer was compromised and used may be liable.)

IT Outsourcers

Most companies outsource at least some of their IT activities, and many companies outsource most of their IT. In such cases, the IT outsourcing vendor will have to work closely with the CSIRT. Beyond that, the IT outsourcer must have good internal incident response capabilities, and expertise in incident handling may be a factor in selecting an outsourcer.

Some firms turn to IT security outsourcers to detect problems in internal systems and to provide expertise and actual work in the case of security incidents. One advantage of using IT security outsourcers is that they have constant experience with handling attacks, while internal CSIRTs may handle only a handful of attacks each year.

In any IT outsourcing, it is important to have outsourcers intimately connected to the CSIRT, including participating in rehearsals for security incidents.

Law Enforcement

Although law enforcement agencies will not actually be part of the CSIRT, their expertise is invaluable if action is to be taken against the attacker. It is important for the CSIRT to understand how to interact with law enforcement agencies and to know what services law enforcement agencies can provide, such as the copying and protection of evidence. Law enforcement officers may even participate in CSIRT rehearsals.

Live Rehearsals

Once the CSIRT is formed, it must engage in live rehearsals that take it through realistic emergencies. During these tests, the CSIRT's members may discover that they are missing representation from an important functional area, that their approaches to recovery are not viable, that they have misunderstandings about authority and even terminology, and so forth. Live rehearsals also build confidence and speed.

Technology Base

The CSIRT will need some technology to do its job. Some of it the firm already has. Some of it must be purchased separately.

Protection Technology

By definition, incidents occur when protection technology (firewalls, server hardening, etc.) break down. However, good protection technology may be able to reduce the severity of an attack. Protections are discussed extensively in other chapters in this encyclopedia.

Intrusion Detection Systems (IDSs)

Technologies can also help in the response phase of incidents. Most obviously, intrusion detection systems can alert firms to attacks. In addition, IDS log files can help the firm analyze what has happened during an attack.

File Integrity Checkers

To respond to an incident involving a server compromise, it is important to be able to know what changes the attacker has made. File integrity checkers periodically create message digests of important system programs. After an attack, message digests are again computed and compared with message digests before the attack to see which programs had been changed. Unfortunately, few firms use file integrity checkers.

Evidence Collection Technology

Evidence collection technology allows a firm to capture evidence during an attack. If evidence is not collected rapidly and carefully, it will become almost impossible to prosecute attackers or even to understand the attack well. One action is to back up the hard drive on the compromised computer. In addition, transient information stored in RAM during an attack may be crucial to understanding the attack and prosecuting attackers. Special purpose software can handle backups and capture this transient information in ways that will preserve evidence for presentation in court if necessary.

The Problem of Communication

One inevitable problem in CSIRTs is communication. During an attack, e-mail may be compromised, and it is likely to be too slow anyway for important functions. The CSIRT organizer must maintain up-to-date lists of office telephone numbers, home telephone numbers, pager numbers, short message service cellular telephone numbers, e-mail addresses, and other contact information for each member of the CSIRT. This contact information changes so rapidly that it must be updated frequently and aggressively.

In the same vein, each CSIRT member must designate an alternate team member if he or she is unavailable. The CSIRT leader needs to maintain up-to-date contact information for these alternates as well. Ideally, alternates should participate in tests.

The Decision to Prosecute

One issue should be considered very carefully before any incident occurs. This is whether to attempt to punish the attacker. For internal employees, this usually is a fairly straightforward decision because termination and lesser punishments are easy and usually legally safe to administer.

For external hackers, however, prosecution is very difficult, especially if the attacker is a minor. Furthermore, pursuing prosecution may open the firm up to negative publicity and consequent losses in customer and investor confidence. Before attacks occur, the company should develop a policy regarding external (and internal) prosecution with the aid of the corporate legal department and senior managers.

One consideration in developing prosecution policies is that successful prosecution will require the CSIRT to take certain actions discussed later in this chapter. Evidence must be carefully protected and preserved. In addition, during the attack, it may be necessary to allow the attack to continue long enough to gather sufficient evidence for conviction instead of cutting off the attacker as soon as possible.

DURING THE ATTACK

Although preparation before a security incident is important, how well the CSIRT responds to actual security incidents is the litmus test for success.

Discovery and Escalation

First, the attack must be discovered. Often, the employee who discovers the attack is a low-level IT employee or an employee in a functional area such as marketing. A key question to ask any firm to assess its state of security is, "Do your employees know the telephone number for reporting security concerns?" A sticker with this number should be on every telephone in the organization. Often, this is the telephone number of the firm's general security office. Delays in reporting suspected security breaches often result in severe damage before the CSIRT can even start its work.

In other cases, the IT security or general IT staff detects the incident. Often, the firm's intrusion detection system sets off an alarm when there is suspicious activity in the system.

In addition, the IDS maintains log files of events relevant to security. During the later analysis phase, the analysis of these log files is crucial in understanding the nature of the incident.

The on-duty IT and IT security staffs typically handle small incidents themselves. However, the on-duty staff manager must have clear guidelines for when to escalate an incident, that is, when to declare the situation sufficiently dangerous or damaging to convene the CSIRT. In particular, the firm must have good guidelines for classifying attacks as minor or major.

Analysis

The first action of the CSIRT must be to analyze the situation. There is a tendency to act precipitously even when there is little information available. A short period of incident analysis is critical for later success.

The first step is to classify the security incident. Typically, there are three types of security incidents—widespread virus/worm outbreaks, denial-of-service attacks, and hacking attacks. Hacking is the access of a protected resource without authorization or in excess of authorization. Hacking, in other words, is breaking into a client, server, switch, router, or other computer.

The second step is to understand the attack. What vulnerability allowed the hacker to break into the computer? What has he or she already done? How sophisticated does the attack appear to be? In the case of virus/worm outbreaks, how is the virus or worm propagating, and how rapidly is it propagating?

Although analysis typically focuses on internal data in computer and intrusion detection system log files, the CSIRT may also contact other organizations, including the Computer Emergency Response Team/Coordination Center, which maintains information about common hacking breaches and virus attacks. There is a good chance that the current attack uses a method that is common in recent attacks, is well understood, and for which good defenses are known.

Containment

The next step is containment, which means stopping the attack. As noted earlier, the simplest approach to containment is to unplug affected computers. While this is effective and may even prove to be necessary, it is a drastic step that can have important implications for the firm as a whole if an affected server is vital to the firm's operations. Effectively, unplugging a computer is a self-administered denial-of-service attack.

In addition, it may be desirable to allow the attacker to continue working so that the attack can be better understood. Containing the attacker too soon may simply keep them out for a few minutes, after which they may be able to get back in using the same exploit they used to gain entry in the first place. In a virus or worm attack, containment before understanding may result in immediate reinfection through the same vulnerability.

If legal prosecution is a goal, furthermore, it may be important to continue gathering evidence.

On the negative side, if hackers are given sufficient time, they can scatter backdoor attack programs throughout a server's directories, read sensitive information, and take steps to make themselves difficult to detect. Consequently, containment should be initiated as soon as possible.

Recovery

Recovery, also called repair, means getting the computer system back to the state it was in before the attack. This means removing programs the attacker placed on the computer and cleaning up any altered data files.

Recovery usually focuses on program files, because hackers and virus/worms often litter the computer with attack programs that must be rooted out. Attackers also Trojanize legitimate system programs by overwriting them with an attack program but keeping the name the same. To recover the system of program files on the computer, there are three main options: repair during continued operation, restoration from backup tapes, and reinstallation from original installation media.

Repair During Continued Operation

If a company has a good way to recognize attack programs, it may be able to make program file repairs while the computer is still operational. However, this presumes that the firm has done such things as keep message digests for all program files so that changes can be detected. This seldom is done.

Recovery from Backup Tapes

It is more common to reinstall all system and other program files from the last clean backup tape. This can be effective, but it is important to know when the attack began or the backup files may themselves be suspect. In addition, the restoration of program files from backup tapes often requires the system to be taken offline at least temporarily.

Reinstallation from Original Installation Media

In the worst case, the computer's software will have to be reinstalled using original software installation media. It is important to have these media readily available for such eventualities. It is also important to document all configuration changes made during and since the last installation. For instance, patches frequently need to be applied to remove known vulnerabilities, and there usually are many company-specific and computer-specific configuration changes made for other reasons. Reinstallation from original media typically takes the computer down for several hours.

Data Files

The CSIRT will face special nightmares if the attack damaged (or may have damaged) data files. While restoration from the last clean backup tape will work for older files, all data entered since the last backup will be lost unless a real-time journal of changes has been maintained (which is rare). Laboriously, online files can be compared to backup files to identify safe files and files that must be looked at carefully.

Protection Against Subsequent Attacks

The last action of any CSIRT during an attack is to protect the firm against subsequent attacks. Obviously, the specific vulnerability that allowed the attack to succeed must be fixed. In addition, firewalls should be tuned to stop such attacks if possible, and intrusion detection systems must get filters to warn of subsequent attacks conducted the way the current breach was created.

Second, and more subtly, the server that was attacked must be especially hardened against reattacks. Hardening involves the application of all patches to the system and the general configuring of the system for maximum security.

Internet hackers, to demonstrate their successes, often invite other hackers to get into your system. For some period of time after an attack, your "famous" server will continue to be attacked by hackers. If you repair the specific exploit used in the first place, hackers will attempt to find other vulnerabilities to prove that they could break into the server in ways that the original hacker could not.

AFTER THE ATTACK

Sanctions

Sanctions are punishments that a firm directs at attackers. Although anger usually makes a firm eager to sanction attackers, it is not always practical to do so.

Punishing Employees

If the attacker is a current employee, the firm must weight the benefits of prosecution against the simple expedient of firing the employee. (For minor incidents, more mild sanctions may be appropriate.) Firing an employee typically can be done with no publicity, especially if the employee agrees to keep quiet or face legal prosecution. In addition, while prosecution is difficult, employees usually can be terminated fairly easily for attacking IT resources, and it usually is legally safe to terminate them (although U.S. state laws differ, international laws vary greatly, and union agreements may make sanctions difficult to apply in some firms).

Prosecuting Attackers

Prosecuting employees, ex-employees, or other outside attackers in courts of law requires a firm to prove that a particular person committed the attack. This can be difficult, especially if the attacker is using a spoofed IP address or attacked indirectly from a computer previously compromised instead of his or her own computer.

In addition, as noted earlier, the attacker must be traced doing a series of actions (individual actions may not convince a jury), and this may require allowing the attack to continue longer than it would if prosecution were not a goal.

Successful prosecution will depend heavily on the quality of evidence collected. As soon as possible, crucial information on the attacked computer must be documented or backed up. Otherwise, the defense will argue that the evidence is tainted and that many things could have caused later changes.

In addition, the CSIRT needs to carefully document how evidence was collected and who collected the evidence. After evidence is collected, the firm must maintain a clear chain-of-evidence documenting who had the evidence at all times and how it was collected and protected after collection. Any breakdown in documentation can ruin a case. Judges often will not even allow a jury to see evidence that does not meet strict legal evidence requirements.

For prosecution, it is critical to train the CSIRT beforehand in evidence collection and preservation. In fact, it is highly desirable to call in local authorities or the FBI. They can make backups of your hard drives and collect other information in a way likely to get a conviction. Forensics is the application of science to criminal investigations. (Kruse and Heiser [2002] said that computer forensics involves the preservation, identification, extraction, documentation, and interpretation of computer media for evidentiary and/or root cause analysis.) This is not a field for amateurs.

Lawsuits

For attacks that produce only modest amounts of damage, the authorities often are reluctant to prosecute. A firm may still be able to act through the legal system by suing the attacker for actual damages and perhaps also for punitive damages. These damages and the attacker's cost of legal defense may be a satisfactory way to exact some measure of revenge. However, lawsuits are expensive and reveal the attack to public. In addition, if the lawsuit is viewed by the hacker community as "lame," the firm may become a popular target of revenge attacks by other hackers.

Reprisal Attacks

One option is to attack the attacker back—essentially, hacking the attacker's computer. This may seem attractive, but it is almost always a very bad idea. First, attackers often use intermediate victim computers to launch attacks. Reprisal attacks are likely to hurt an innocent victim. (If a firm feels that companies that allow their computers to be hacked are not innocent, it should keep in mind that it has just been hacked itself.)

Of course, more fundamentally, reprisal attacks are almost always illegal. The jailing of the CSIRT is not a goal of incident response. Also, a CSIRT is an official organization within the firm; liability for a reprisal attack is likely to extend to senior officers and even the board of directors.

Postmortem Analysis

After things have stabilized, it is important to conduct a postmortem analysis to review what happened during the incident. The goal is to improve future responses to incidents. Things that worked well should be noted, but problems also should be highlighted and fixes created for the future. After the exhaustion of an incident and the need to catch up on work postponed during the incident, it is difficult to motivate the CSIRT team members to undertake a postmortem analysis. However, given the valuable information gained during the actual incident, as opposed to simple live rehearsals, the information gathered in the postmortem analysis can produce strong benefits.

There should also be a follow-up analysis one to six months later to determine which recommendations made during a postmortem analysis have been implemented and which have not.

CONCLUSION

When major IT security incidents happen, companies cannot sit back and wonder what they should do. Before the attack, companies need to organize and train cross-disciplinary computer security incident response teams —also called computer emergency response teams. Training using realistic rehearsals is very important for ensuring adequate and rapid CSIRT responses during an emergency.

During an attack, once an incident is discovered and escalated by the on-duty staff to a major incident, the CSIRT is activated. The CSIRT analyzes the attack so that appropriate responses can be made, contains the attack to prevent further damage, recovers program files and perhaps data files, gathers information needed for internal punishment or legal prosecution, hardens the system to prevent reattacks, and conducts a postmortem analysis to refine its incident response abilities.

GLOSSARY

Analysis Incident response phase just after discovery phase. The CSIRT determines how the attack was made and what was done to affected systems.

Breach Another name for a security incident.

Computer emergency response team Another name for computer security incident response team.

Computer security incident response team Cross-disciplinary team created to respond to major security incidents.

Containment Phase in incident response in which the damage to systems is stopped.

Discovery Phase in incident response in which a security incident is first discovered and reported.

Escalation The act of determining that a particular security incident is a major breach and that the CSIRT must be activated.

Forensics The application of science to criminal investigations.

Hacking Taking over a computer. Using a computer without authorization or in excess of authorization.

Hardening Removing vulnerabilities from a host computer and configuring the computer for high security.

Incident Another name for a security incident.

Incident response The multiphase process of responding to a security incident.

Intrusion detection system Tool that signals that an attack appears to be underway and that collects data in log files for analysis during and after an incident.

IT Information technology, computer hardware, computer software, networking, and data.

IT security Security of the IT function, as opposed to physical and building security.

IT security staff The IT security staff given responsibility for IT security, as opposed to physical and building security.

Legal department Department containing the firm's legal staff.

Outsourcer External company that provides IT services to the firm.

Postmortem analysis Assessment conducted after a security incident to determine how to better respond in the future.

Prosecution The attempt to convict an attacker in a court of law.

Public relations Department in a firm charged with communicating with the public and the media on behalf of the firm.

Punishment Taking action against an attacker; both sanctioning employee attackers and prosecuting and suing external attackers.

Repair Phase in incident response in which the damaged system is brought back to correct operation.

Reprisal attack Attacking the attacker (usually illegal).

Security breach Another name for a security incident.

Security incident A virus attack, denial-of-service attack, or hack (computer break-in).

Trojanize Replace a legitimate file by an attacker's file, giving the attacker's file the same name as the legitimate file.

CROSS REFERENCES

See *Computer Viruses and Worms; Denial of Service Attacks; Disaster Recovery Planning; Guidelines for a Comprehensive Security System; Intrusion Detection Techniques; Physical Security.*

REFERENCES

Kruse, W. G. II, & Heiser, J. G. (2002). *Computer forensics: Incident response essentials.* Boston: Addison-Wesley.

Panko, R. R. (2004). *Corporate computer and network security.* Upper Saddle River, NJ: Prentice-Hall.

FURTHER READING

Mandia, K., & Prosise, C. (2001). *Incident response.* New York: Osborne/McGrawHill.

Northcutt, S., & Novak, J. (2001). *Network intrusion detection: An analyst's handbook.* Indianapolis: New Riders.

Schultz, E. E., & Shumway, R. (2002). *Incident response.* Indianapolis: New Riders.

Computer Viruses and Worms

Robert Slade, *Consultant*

INTRODUCTION

Computer viruses are unique among the many security problems in the fact that someone else being infected increases the risk to you. However, viruses also seem to be surrounded by myths and misunderstandings. It is hoped that this chapter will help to set the record straight.

History of Computer Viruses and Worms

Many claims have been made for the existence of viruses prior to the 1980s, but, so far, these claims have not been accompanied by proof. The Core Wars programming contests did involve self-replicating code, but usually within a structured and artificial environment.

The general perception of computer viruses, even among security professionals, has concentrated on their existence in personal computers, and particularly "Wintel" (Windows/Intel)-type systems. This is despite the fact that Fred Cohen's seminal academic work took place on mainframe and minicomputers in the mid-1980s. The first e-mail virus was spread in 1987. The first virus hoax message (then termed a "metavirus") was proposed in 1988. Even so, virus and malware research has been neglected, possibly because malware does not fit easily into the traditional access control security models.

At least two Apple II viruses are known to have been created in the early 1980s. However, it was not until the end of the decade (and 1987 in particular) that knowledge of real viruses became widespread, even among security experts. For many years boot sector infectors and file infectors were the only types of common viruses. These programs spread relatively slowly, primarily distributed on floppy disks, and were thus slow to disseminate geographically. However, these viruses tended to be very long-lived.

During the early 1990s virus writers started experimenting with various functions intended to defeat detection. (Some forms had seen limited trials earlier.) Among these were polymorphism, to change form in order to defeat scanners, and stealth, to attempt to confound any type of detection. None of these virus technologies had a significant impact. Most viruses using these "advanced" technologies were easy to detect because of a necessary increase in program size.

Although demonstration programs had been created earlier, the mid-1990s saw the introduction of macro and script viruses in the wild. These were initially confined to word processing files, particularly files associated with the Microsoft Office Suite. However, the inclusion of programming capabilities eventually led to script viruses in many objects that would normally be considered to contain data only, such as Excel spreadsheets, PowerPoint presentation files, and e-mail messages. This fact led to greatly increased demands for computer resources among antiviral systems, since many more objects had to be tested, and Windows OLE (object linking and embedding) format data files presented substantial complexity to scanners. Macro viruses also increase in new variant forms very quickly, since the virus carries its own source code, and anyone who obtains a copy can generally modify it and create a new member of the virus family.

E-mail viruses became the major new form in the late 1990s and early 2000s. These viruses may use macro capabilities, scripting, or executable attachments to create e-mail messages or attachments sent out to e-mail addresses harvested from the infected machine. E-mail viruses spread with extreme rapidity, distributing themselves worldwide in a matter of hours. Some versions create so many copies of themselves that corporate and even service provider mail servers are flooded and cease to function. E-mail viruses are very visible and so tend to be identified within a short space of time, but many are macros or scripts and so generate many variants.

With the strong integration of the Microsoft Windows operating system with its Internet Explorer browser, Outlook mailer, Office suite, and system scripting, recent viruses have started to blur the normal distinctions. A document sent as an e-mail file attachment can make a call to a Web site that starts active content that installs a remote access tool acting as a portal for the client portion of a distributed-denial-of-service network. Indeed, not only are viruses starting to show characteristics that are similar to each other, but functions from completely different types of malware are beginning to be found together in the same programs, leading to a type of malware convergence.

Recently, many security specialists have stated that the virus threat is reducing, since, despite the total number of virus infections being seen, the prevalent viruses are now almost universally e-mail viruses and therefore constitute a single threat with a single fix. This ignores the fact that although almost all major viruses now use e-mail as a distribution and reproduction mechanism, there are a great many variations in the way e-mail is used. For example, many viruses use Microsoft's Outlook mailer to spread, and reproduction can be prevented simply by removing Outlook from the system. However, other viruses may make direct calls to the Mail Application Programing Interface (MAPI), which is used by a number of mail user programs, while others carry the code for mail server functions within their own body. A number of e-mail viruses distribute themselves to e-mail addresses found in the Microsoft Outlook address book files, while others may harvest addresses from anywhere on the computer hard drive, or may actually take control of the Internet network connection and collect contact data from any source viewed online.

Because the work has had to deal with detailed analysis of low-level code, virus research has led to significant advances in the field of forensic programming. However, to date computer forensic work has concentrated on file recovery and decryption, so the contributions in this area likely still lie in the future.

Many computer pundits, as well as some security experts, have proposed that computer viruses are a result of the fact that currently popular desktop operating systems have only nominal security provisions. They further suggest that viruses will disappear as security functions are added to operating systems. This thesis ignores the fact, well established by Cohen's research and subsequently confirmed, that viruses use the most basic of computer functions, and that a perfect defense against viruses is impossible. This is not to say that an increase in security measures by operating system vendors could not reduce the risk of viruses: the current danger could be drastically reduced with relatively minor modifications to system functions.

It is going too far to say (as some have) that the very existence of viral programs, and the fact that both viral strains and the numbers of individual infections are growing, means that computers are finished. At the present time, the general public is not well informed about the virus threat, and so more copies of viral programs are being produced than are being destroyed.

Indeed, no less an authority than Fred Cohen has championed the idea that viral programs can be used to great effect. An application using a viral form can improve performance in the same way that computer hardware benefits from parallel processors. It is, however, unlikely that viral programs can operate effectively and usefully in the current computer environment without substantial protective measures being built into them. A number of virus and worm programs have been written with the obvious intent of proving that viruses could carry a useful payload, and some have even had a payload that could be said to enhance security. Unfortunately, all such viruses have created serious problems themselves.

Virus Definition

A computer virus is a program written with functions and intent to copy and disperse itself without the knowledge and cooperation of the owner or user of the computer. A final definition has not yet been agreed upon by all researchers. A common definition is "a program which modifies other programs to contain a possibly altered version of itself." This definition is generally attributed to Fred Cohen from his seminal research in the mid-1980s, although Dr. Cohen's actual definition is in mathematical form. Another possible definition is an entity that uses the resources of the host (system or computer) to reproduce itself and spread, without informed operator action.

Cohen is generally held to have defined the term "computer virus" in his thesis (published in 1984). (The suggestion for the use of the term virus is credited to Len Adleman, his seminar advisor.) However, his original definition covers only those sections of code that, when active, attach themselves to other programs. This, however, neglects many of the programs that have been most successful "in the wild." Many researchers still insist on Cohen's definition and use other terms such as "worm" and "bacterium" for those viral programs that do not attack programs. Currently, viruses are generally held to attach themselves to some object, although the object may be a program, disk, document, e-mail message, computer system, or other information entity.

Computer viral programs are not a "natural" occurrence. Viruses are programs written by programmers. They do not just appear through some kind of electronic evolution. Viral programs are written, deliberately, by people. However, the definition of "program" may include many items not normally thought of in terms of programming, such as disk boot sectors, and Microsoft Office documents or data files that also contain macro programming.

Many people have the impression that anything that goes wrong with a computer is caused by a virus. From hardware failures to errors in use, everything is blamed on a virus. A virus is not just any damaging condition. Similarly, it is now popularly believed that any program that may do damage to your data or your access to computing resources is a virus. Viral programs are not simply programs that do damage. Indeed, viral programs are not always damaging, at least not in the sense of being deliberately designed to erase data or disrupt operations. Most viral programs seem to have been designed to be a kind of electronic graffiti: intended to make the writer's mark in the world, if not his or her name. In some cases a name is displayed, on occasion an address, phone number, company name, or political party.

Is My Computer Infected? What Should I Do?

In many books and articles you will find lists of symptoms to watch for in order to determine whether your computer is infected. These signs include things like running out of memory space, running out of disk space, the computer operating slower than normal, files changing size, and so forth. In fact, many factors will create these same effects, and current viruses seldom do. The best way to determine whether you have been infected by a virus is to get and use an antiviral scanner. In fact, get more than one. With the rapid generation of new viruses these days, it is quite possible for a maker of antivirus software to make mistakes in updating signatures. Therefore, having a second check on a suspected virus is always a good idea.

One scanner for the Wintel platform is F-PROT. It is available in a DOS version, free of charge from http://www.f-secure.com. (Look under "Downloads" and "Tools.") Although a DOS scanner has some limitations in a Windows envrionment, it will still be able to identify infected files, and is quite good at picking out virus infections within e-mail system files (the files of messages held on your computer).

Another scanner available for free is AVG, from http://www.grisoft.com. This one also does a good job of scanning, and will even update itself automatically (although that feature is problematic on some machines).

The various commercial antiviral producers generally produce trial versions of their software, usually limited in some way.

In regard to the suggestion to use more than one scanner, it should be noted that a number of successful software publishers have included functions that conflict with software from other vendors. These products should be avoided, because of the previously noted possibility of failure in a single protection program. The use of free software, and the purchase of software from companies that provide such free versions, is recommended, since the existence of these free scanners, and their use by other people, actually reduces your risk, since there will be fewer instances of viruses reproducing and trying to spread.

Readers may be surprised at the recommendation to use free software: there is a general assumption that commercial software must be superior to that provided free of charge. It should be noted that the author of this chapter is a specialist in the evaluation of security and, particularly, antiviral software. The reader can be assured that it can be proven that, in the case of antiviral software, free software is as effective, and in some cases superior to, many very expensive products.

TROJAN HORSES, VIRUSES, WORMS, RATS, AND OTHER BEASTS

Malware is a relatively new term in the security field. It was created to address the need to discuss software or programs intentionally designed to include functions for penetrating a system, breaking security policies, or carrying malicious or damaging payloads. Since this type of software has started to develop a bewildering variety of forms—such as backdoors, data diddlers, DDoS (distributed denial of service), hoax warnings, logic bombs, pranks, RATs (remote access Trojans), Trojans, viruses, worms, and zombies—the term malware has come to be used for the collective class of *mal*icious soft*ware*. The term is, however, often used very loosely simply as a synonym for virus, in the same way that virus is often used simply as a description of any type of computer problem. This chapter will attempt to define the problem more accurately, and to describe the various types of malware.

Viruses are the largest class of malware, both in terms of numbers of known entities and in impact in the current computing environment. Viruses, therefore, tend to be synonymous, in the public mind, with all forms of malware.

Programming bugs or errors are generally not included in the definition of malware, although it is sometimes difficult to make a hard and fast distinction between malware and bugs. For example, if a programmer left a buffer overflow in a system and it creates a loophole that can be used as a backdoor or a maintenance hook, did he do it deliberately? This question cannot be answered technically, although we might be able to guess at it, given the relative ease of use of a given vulnerability.

In addition, it should be noted that malware is not just a collection of utilities for the attacker. Once launched, malware can continue an attack without reference to the author or user, and in some cases will expand the attack to other systems. There is a qualitative difference between malware and the attack tools, kits, or scripts that must operate under an attacker's control and those not considered to fall within the definition of malware. There are gray areas in this aspect as well, since RATs and DDoS zombies provide unattended access to systems, but need to be commanded in order to deliver a payload.

Trojans

Trojans, or Trojan horse programs, are the largest class of malware. However, the term is subject to much confusion, particularly in relation to computer viruses.

A Trojan is a program that pretends to do one thing while performing another, unwanted action. The extent of the "pretense" may vary greatly. Many of the early PC Trojans relied merely on the filename and a description on a bulletin board. "Log-in" Trojans, popular among university student mainframe users, mimicked the screen display and the prompts of the normal log-in program and could, in fact, pass the user name and password along to the valid log-in program at the same time as they stole the user data. Some Trojans may contain actual code that does what it is supposed to be doing while performing additional nasty acts that it does not tell you about.

An additional confusion with viruses involves Trojan horse programs that may be spread by e-mail. In years past, a Trojan program had to be posted on an electronic bulletin board system or a file archive site. Because of the static posting, a malicious program would soon be identified and eliminated. More recently, Trojan programs have been distributed by mass e-mail campaigns, by posting on Usenet newsgroup discussion groups, or through automated distribution agents (bots) on Internet relay chat (IRC) channels. Since source identification in these communications channels can be easily hidden, Trojan programs can be redistributed in a number of disguises, and specific identification of a malicious program has become much more difficult.

Some data security writers consider that a virus is simply a specific example of the class of Trojan horse programs. There is some validity to this usage since a virus is an unknown quantity that is hidden and transmitted along with a legitimate disk or program, and any program can be turned into a Trojan by infecting it with a virus. However, the term virus more properly refers to the added, infectious code rather than the virus/target combination. Therefore, the term trojan refers to a deliberately misleading or modified program that does not reproduce itself.

A major aspect of Trojan design is the social engineering (fraudulent or deceptive) component. Trojan programs are advertised (in some sense) as having a positive component. The term positive can be in dispute, since a great many Trojans promise pornography or access to pornography, and this still seems to be depressingly effective. However, other promises can be made as well. A recent e-mail virus, in generating its messages, carried a list of a huge variety of subject lines, promising pornography, humor, virus information, an antivirus program, and information about abuse of the recipient's e-mail account. Sometimes the message is simply vague and relies on curiosity.

Social engineering really is nothing more than a fancy name for the type of fraud and confidence games that have existed since snakes started selling apples. Security types tend to prefer a more academic sounding definition, such as the use of nontechnical means to circumvent security policies and procedures. Social engineering can range from simple lying (such as a false description of the function of a file) to bullying and intimidation (in order to pressure a low-level employee into disclosing information), to association with a trusted source (such as the user name from an infected machine).

Worms

A worm reproduces and spreads, like a virus, and unlike other forms of malware. Worms are distinct from viruses, although they may have similar results. Most simply, a worm may be thought of as a virus with the capacity to propagate independently of user action. In other words, they do not rely on (usually) human-initiated transfer of data between systems for propagation, but instead spread across networks of their own accord, primarily by exploiting known vulnerabilities in common software.

Originally, the distinction was made that worms used networks and communications links to spread, and that a worm, unlike a virus, did not directly attach to an executable file. In early research into computer viruses, the terms worm and virus tended to be used synonymously, it being felt that the technical distinction was unimportant to most users. The technical origin of the term "worm program" matched that of modern distributed processing experiments: a program with "segments" working on different computers, all communicating over a network (Shoch & Hupp, 1982).

The first worm to garner significant attention was the Internet Worm of 1988, to be discussed in detail in the section Worms (First and Third Generations). Recently, many of the most prolific virus infections have not been strictly viruses but have used a combination of viral and worm techniques to spread more rapidly and effectively. LoveLetter was an example of this convergence of reproductive technologies. Although infected e-mail attachments were perhaps the most widely publicized vector of infection, LoveLetter also spread by actively scanning attached network drives, infecting a variety of common file types. This convergence of technologies will be an increasing problem in the future. Code Red and a number of Linux programs (such as Lion) are modern examples of worms. (Nimda is an example of a worm, but it also spreads in a number of other ways, so it could be considered to be an e-mail virus and multipartite as well.)

Viruses

A virus is defined by its ability to reproduce and spread. A virus is not just anything that goes wrong with a computer, and virus is not simply another name for malware. Trojan horse programs and logic bombs do not reproduce themselves.

A worm, which is sometimes seen as a specialized type of virus, is currently distinguished from a virus because a virus generally requires an action on the part of the user to trigger or aid reproduction and spread. The action on the part of the user is generally a common function, and the user generally does not realize the danger of the action or the fact that he or she is assisting the virus.

The only requirement that defines a program as a virus is that it reproduces. There is no necessity that the virus carries a payload, although a number of viruses do. In many cases (in most cases of "successful" viruses), the payload is limited to some kind of message. A deliberately damaging payload, such as erasure of the disk, or system files, usually restricts the ability of the virus to spread because the virus uses the resources of the host system. In some cases, a virus may carry a logic bomb or time bomb which triggers a damaging payload on a certain date or under a specific, often delayed, condition.

Because a virus spreads and uses the resources of the host, it affords a kind of power to software that parallel processors provide to hardware. Therefore, some have theorized that viral programs could be used for beneficial purposes, similar to the experiments in distributed processing that are testing the limits of cryptographic strength. (Various types of network management functions, and updating of system software, are seen as candidates.) However, the fact that viruses change systems and applications is seen as problematic in its own right. Many viruses that carry no overtly damaging payload still create problems with systems. A number of virus and worm programs have been written with the obvious intent of proving that viruses could carry a useful payload, and some have even had a payload that could be said to enhance security. Unfortunately, all such viruses have created serious problems themselves. The difficulties of controlling viral programs have been addressed in theory, but the solutions are also known to have faults and loopholes.

Logic Bombs

Logic bombs are software modules set up to run in a quiescent state but to monitor for a specific condition, or set of conditions, and to activate their payload under those conditions. A logic bomb is generally implanted in or coded as part of an application under development or maintenance. Unlike a RAT or Trojan it is difficult to implant a logic bomb after the fact. There are numerous examples of this type of activity, usually based upon actions taken by a programmer to deprive a company of needed resources if employment was terminated.

A Trojan or a virus may contain a logic bomb as part of the payload. A logic bomb involves no reproduction, no social engineering.

A persistent legend in regard to logic bombs involves what is known as the salami scam. According to the story, this involves the siphoning off of small amounts on money (in some versions, fractions of a cent) credited to the account of the programmer, over a very large number of transactions. Despite the fact that these stories appear in a number of computer security texts, the author has a standing challenge to anyone to come up with a documented case of such a scam. Over a period of eight years, the closest anyone has come is a story about a fast food clerk who diddled the display on a drive through window, and collected an extra dime or quarter from most customers.

Other Related Terms

Hoax virus warnings or alerts have an odd double relation to viruses. First off, hoaxes are usually warnings about “new” viruses: new viruses that do not, of course, exist. Second, hoaxes generally carry a directive to the user to forward the warning to all addresses available to them. Thus these descendants of chain letters form a kind of self-perpetuating spam.

Hoaxes use an odd kind of social engineering, relying on people’s naturally gregarious nature and desire to communicate, and on a sense of urgency and importance, using the ambition that people have to be the first to provide important new information.

Hoaxes do, however, have common characteristics that can be used to determine whether their warnings may be valid:

- Hoaxes generally ask the reader to forward the message.
- Hoaxes make reference to false authorities such as Microsoft, AOL, IBM, and the FCC (none of which issue virus alerts), or to completely false entities.
- Hoaxes do not give specific information about the individual or office responsible for analyzing the virus or issuing the alert;
- Hoaxes generally state that the new virus is unknown to authorities or researchers.
- Hoaxes often state that there is no means of detecting or removing the virus.
- Many of the original hoax warnings stated only that you should not open a message with a certain phrase in the subject line. (The warning, of course, usually contained that phrase in the subject line. Subject line filtering is known to be a very poor method of detecting malware.)
- Hoaxes often state that the virus does tremendous damage, and is incredibly virulent.
- Hoax warnings very often contain A LOT OF CAPITAL LETTER SHOUTING AND EXCLAMATION MARKS!!!!!!!!!!
- Hoaxes often contain technical sounding nonsense (technobabble), such as references to nonexistent technologies like “nth complexity binary loops.”

It is wisest, in the current environment, to doubt all virus warnings, unless they come from a known and historically accurate source, such as a vendor with a proven record of providing reliable and accurate virus alert information, or preferably an independent researcher or group. It is best to check *any* warnings received against known virus encyclopedia sites. It is best to check more than one such site: in the initial phases of a “fast burner” attack some sites may not have had time to analyze samples to their own satisfaction, and the better sites will not post information they are not sure about.

Remote access Trojans (RATs) are programs designed to be installed, usually remotely, after systems are installed and working (and not in development, as is the case with logic bombs and backdoors). Their authors would generally like to have the programs referred to as remote administration tools, in order to convey a sense of legitimacy.

When a RAT program has been run on a computer, it will install itself in such a way as to be active every time the computer is started subsequent to the installation. Information is sent back to the controlling computer (sometimes via an anonymous channel such as IRC), noting that the system is active. The user of the command computer is now able to explore the target, escalate access to other resources, and install other software, such as DDoS zombies, if so desired.

DDoS (distributed denial of service) is a modified denial of service (DoS) attack. Denial of service attacks do not attempt to destroy or corrupt data, but attempt to use up a computing resource to the point where normal work cannot proceed. The structure of a DDoS attack requires a master computer to control the attack, a target of the attack, and a number of computers in the middle that the master computer uses to generate the attack. These computers in between the master and the target are variously called agents or clients, but are usually referred to as running “zombie” programs. The existence of a large number of agent computers in a DDoS attack acts to multiply the effect of the attack, and also helps to hide the identity of the originator of the attack.

There is a lot of controversy over a number of technologies generally described as adware or spyware. Most people would agree that the marketing functions are not specifically malicious, but what one person sees as "aggressive selling" another will see as an intrusion or invasion of privacy.

Shareware or freeware programs may have advertising for a commercial version or a related product, and users may be asked to provide some personal information for registration or to download the product. For example, an unregistered copy of the WinZip archiving program typically asks the user to register when the program is started, and the free version of the QuickTime video player asks the user to buy the commercial version every time the software is invoked. Adware, however, is generally a separate program installed at the same time as a given utility or package, and continues to advertise products, even when the desired program is not running. Spyware is, again, a system distinct from the software the user installed, and passes more information than simply a user name and address back to the vendor: often these packages will report on Web sites visited or other software installed on the computer, and possibly compile detailed inventories of the interests and activities of a user.

The discussion of spyware often makes reference to cookies or Web bugs. Cookies are small pieces of information in regard to persistent transactions between the user and a Web site, but the information can be greater than the user realizes, for example, in the case of a company that provides content, such as banner ads, to a large number of sites. Cookies, limited to text data, are not malware, and can have no executable malicious content. Web bugs are links on a Web page or embedded in e-mail messages that contain links to different Web sites. A Web bug therefore passes a call, and information, unknown to the user, to a remote site. Most commonly a Web bug is either invisible or unnoticeable (typically it is one pixel in size) in order not to alert the user to its presence. (There is a persistent chain letter hoax that tells people to forward the message because it is part of a test of an e-mail tracking system. Although all such reports to date are false, if such a system were to be implemented, Web bugs would likely be the technology used. Even then the system would not be reliable: Web bugs in e-mail rely on an e-mail system calling a Web browser function, and although this typically happens automatically with systems like Microsoft's Outlook and Internet Explorer, a mailer like Pegasus requires the function call to be established by the user, and warns the user when it is being invoked.

Pranks are very much a part of the computer culture. So much so that you can now buy commercially produced joke packages that allow you to perform "Stupid Mac (or PC, or Windows) Tricks." There are numberless pranks available as shareware. Some make the computer appear to insult the user; some use sound effects or voices; some use special visual effects. A fairly common thread running through most pranks is that the computer is, in some way, nonfunctional. Many pretend to have detected some kind of fault in the computer (and some pretend to rectify such faults, of course making things worse). One entry in the virus field is PARASCAN, the paranoid scanner. It pretends to find large numbers of infected files, although it doesn't actually check for any infections.

Generally speaking, pranks that create some kind of announcement are not malware: viruses that generate a screen or audio display are actually quite rare. The distinction between jokes and Trojans is harder to make, but pranks are intended for amusement. Joke programs may, of course, result in a denial of service if people find the prank message frightening.

One specific type of joke is the "easter egg," a function hidden in a program, and generally accessible only by some arcane sequence of commands. These may be seen as harmless, but note that they do consume resources, even if only disk space, and also make the task of ensuring program integrity very much more difficult.

FIRST GENERATION VIRUSES

Many of the books and articles currently available to explain about viruses were written based on the research that was done on the first generation of viruses. While these programs are still of interest to those who study the internal structures of operating systems, they operated much differently than the current crop of malicious software. First generation viruses tended to spread very slowly but hang around in the environment for a long time. Later viruses tended to spread very rapidly, but also to die out relatively quickly.

Boot Sector Viruses

Viruses are generally partly classified by the objects to which they attach. Worms may be seen as a type of virus that attaches to nothing. Most desktop computer operating systems have some form of boot sector, a specific location on disk that contains programming to bootstrap the startup of a computer. Boot sector infectors (BSIs) replace or redirect this programming in order to have the virus invoked, usually as the first programming running on the computer.

Boot sector infectors would not appear to fit the definition of a virus infecting another program, since BSIs can be spread by disks that do not contain any program files. However, the boot sector of a normal MS-DOS disk, whether or not it is a "system" or bootable disk, always contains a program (even if it only states that the disk is not bootable), and so it can be said that a BSI is a "true" virus.

The terminology of BSIs comes from MS-DOS systems, and this leads to some additional confusion. The physical "first sector" on a hard drive is not the operating system boot sector. On a hard drive the boot sector is the first "logical" sector. The number one position on a hard drive is the master boot record (MBR). Some viral programs, such as the Stoned virus, always attack the physical first sector: the boot sector on floppy disks and the master boot record on hard disks. Thus viral programs that always attack the boot sector might be termed "pure" BSIs, whereas programs like Stoned might be referred to as an "MBR type" of BSI. The term boot sector infector is used for all of them, though, since all of them infect the boot sector on floppy disks.

File-Infecting Viruses

A file infector infects program (object) files. System infectors that infect operating system program files (such as COMMAND.COM in DOS) are also file infectors. File infectors can attach to the front of the object file (prependers), attach to the back of the file and create a jump at the front of the file to the virus code (appenders), or overwrite the file or portions of it (overwriters). A classic is Jerusalem. A bug in early versions caused it to add itself over and over again to files, making the increase in file length detectable. (This has given rise to the persistent myth that it is a characteristic of a virus that it will fill up all disk space eventually: by far the majority of file infectors add minimally to file lengths.)

Polymorphic Viruses

Polymorphism (literally many forms) refers to a number of techniques that attempt to change the code string on each generation of a virus. These vary from using modules that can be rearranged to encrypting the virus code itself, leaving only a stub of code that can decrypt the body of the virus program when invoked. Polymorphism is sometimes also known as self-encryption or self-garbling, but these terms are imprecise and not recommended. Examples of viruses using polymorphism are Whale and Tremor. Many polymorphic viruses use standard "mutation engines" such as MtE. These pieces of code actually aid detection because they have a known signature.

A number of viruses also demonstrate some form of active detection avoidance, which may range from disabling of on-access scanners in memory to deletion of antivirus and other security software (Zonealarm is a favorite target) from the disk.

Virus Creation Kits

The term "kit" usually refers to a program used to produce a virus from a menu or a list of characteristics. Use of a virus kit involves no skill on the part of the user. Fortunately, most virus kits produce easily identifiable code. Packages of antiviral utilities are sometimes referred to as tool kits, occasionally leading to confusion of the terms.

MACRO VIRUSES (SECOND GENERATION)

A macro virus uses macro programming of an application such as a word processor. (Most known macro viruses use Visual Basic for Applications in Microsoft Word: some are able to cross between applications and function in, for example, a PowerPoint presentation and a Word document, but this ability is rare.) Macro viruses infect data files, and tend to remain resident in the application itself by infecting a configuration template such as MS Word's normal.dot. Although macro viruses infect data files, they are not generally considered to be file infectors: a distinction is generally made between program and data files. Macro viruses can operate across hardware or operating system platforms as long as the required application platform is present. (For example, many MS Word macro viruses can operate on both the Windows and Macintosh versions of MS Word.) Examples are Concept and CAP. Melissa is also a macro virus, in addition to being an e-mail virus: it mailed itself around as an infected document.

What Is a Macro?

As noted above, a macro is a small piece of programming contained in a larger data file. This differentiates it from a script virus, which is usually a standalone file that can be executed by an interpreter, such as Microsoft's Windows Script Host (.vbs files). A script virus file can be seen as a data file in that it is generally a simple text file, but it usually does not contain other data, and generally has some indicator (such as the .vbs extension) that it is executable. LoveLetter is a script virus.

Free Technology to Avoid Macro Viruses

A recommended defence is MacroList, written by A. Padgett Peterson. This is a macro itself, available for both Wintel and Macintosh machines. It will list all the macros in a document. Since most documents should not contain macros, any document that does should either have a really good reason for it, or be looked at with suspicion. You can find MacroList at http://www2.gdi.net\~padgett\index.htm.

E-MAIL VIRUSES (THIRD GENERATION)

With the addition of programmable functions to a standard e-mail user agent (usually Microsoft's Outlook), it became possible for viruses to spread worldwide in mere hours, as opposed to months.

The "Start" of E-mail Viruses: Melissa

Melissa was far from the first e-mail virus. The first e-mail virus to successfully spread in the wild was the Christma exec, in the fall of 1987. However, Melissa was certainly the first of the "fast burner" e-mail viruses, and the first to come to wide public attention.

The virus, generally referred to as W97M.Melissa, is a Microsoft Word macro virus. The name "Melissa" comes from the class module that contains the virus. The name is also used in the registry flag set by the virus.

The virus is spread, of course, by infected Word documents. What has made it the "bug du jour" is that it spreads *itself* via e-mail.

Melissa was originally posted to the alt.sex newsgroup. At that time it was list.doc, and purported to be a list of passwords for sex sites.

If you get a message with a Melissa infected document and do whatever you need to do to "invoke" the attachment, and have Word on your system as the default program for .doc files, Word starts up, reads in the document, and the macro is ready to start.

Assuming that the macro starts executing, several things happen.

The virus first checks to see whether Word 97 (Word 8) or Word 2000 (Word 9) is running. If so, it reduces the level of the security warnings on Word so that you will receive no future warnings. In Word 97, the virus disables the Tools/Macro menu commands, the Confirm Conversions option, the MS Word macro virus protection, and the Save Normal Template prompt. It "upconverts" to Word 2000 quite nicely and there disables the Tools/Macro/Security menu.

Specifically, under Word 97 it blocks access to the Tools|Macro menu item, meaning you cannot check any macros. It also turns off the warnings for conversion, macro detection, and saving modifications to the normal.dot file. Under Word 2000 it blocks access to the menu item that allows you to raise your security level and sets your macro virus detection to the lowest level, that is, none. Since the access to the macro security menu item is blocked, you must delete the infected normal.dot file in order to regain control of your security settings. Note that this will also lose all of your global templates and macros. Word users who make extensive use of macros are advised to keep a separate backup copy of a clean normal.dot in some safe location to avoid problems with macro virus infections.

After this, the virus checks for the HKEY_CURRENT_USER\Software\Microsoft\Office\Melissa?\ registry key with a value of "... by Kwyjibo." (The "kwyjibo" entry seems to be a reference to the "Bart the Genius" episode of the "Simpsons" television program where this word was used to win a Scrabble match.)

If this is the first time you have been infected, the macro starts up Outlook 98 or higher, in the background, and sends itself as an attachment to the "top" 50 names in *each* of your address lists. (Melissa will not use Outlook Express. Also, Outlook 97 will not work.) Most people have only one (the default is "Contacts"), but if you have more than one then Outlook will send more than 50 copies of the message. Outlook also sorts address lists such that mailing lists are at the top of the list, so this can get a much wider dispersal than just 50 copies of the message/virus.

Once the messages have been sent, the virus sets the Melissa flag in the registry, and looks for it to check whether to send itself out on subsequent infections. If the flag does not persist, then there will be subsequent mass mailings. Because the key is set in HKEY_CURRENT_USER, system administrators may have set permissions such that changes made are not saved, and thus the key will not persist. In addition, multiple users on the same machine will likely each trigger a separate mailout, and the probability of cross infection on a common machine is very high.

Since it is a macro virus, it will infect your normal.dot, and will infect all documents thereafter. The macro within normal.dot is "Document_Close()" so that any document that is worked on (or created) will be infected when it is closed. When a document is infected the macro inserted is "Document_Open()" so that the macro runs when the document is opened.

Note that not using Outlook does not protect you from the virus, it only means that the 50 copies will not be automatically sent out. If you use Word but not Outlook, you will still be infected and may still send out infected documents on your own. Originally the virus would not invoke the mailout on Macintosh systems. However, infected documents would be stored, and, recently, when Outlook became available for Macs, there was a second wave of Melissa mailings.

The message appears to come from the person just infected, of course, since it really is sent from that machine. This means that when you get an "infected" message it will probably appear to come from someone you know and deal with. The subject line is "Important Message From: [name of sender]" with the name taken from the registration settings in Word. The text of the body states "Here is that document you asked for... don't show anyone else;-)." Thus, the message is easily identifiable: that subject line, the very brief message, and an attached Word document (file with a .doc extension to the filename).

However, note that, as with any Microsoft Word macro virus, the source code travels with the infection, and it was very easy for people to create variations of Melissa. Within days of Melissa there was a similar Excel macro virus, called "Papa."

One rather important point: the document passed is the active document, not necessarily the original posted on alt.sex. So, for example, if I am infected, and prepare some confidential information for you in Word, and send you an attachment with the Word document, containing sensitive information that neither you nor I want made public and you read it in Word, and you have Outlook on your machine, then that document will be mailed out to the top 50 people in your address book, and so forth.

How to Avoid E-mail Viruses

It really is very simple to avoid e-mail viruses: don't double-click on any attachments that come with your e-mail. We used to say not to run any programs that came from someone you don't know, but many e-mail viruses spread using the identity of the owner of the infected computer, so that is no longer any protection. Do not run anything you receive, unless you know, from some separate verification, that this person intended to send you something, and that it is something you need, and that the person sending it is capable of protecting themselves from infection.

It is also somewhat safer to use a mail program other than Outlook, since some versions of Outlook allowed attachments to run even before the user read the message to which they were attached.

WORMS (FIRST AND THIRD GENERATIONS)

In autumn 1988, the Internet/UNIX/Morris Worm did not actually bring the Internet in general and e-mail in particular to the proverbial grinding halt. It was able to run and propagate only on machines running specific versions of the UNIX operating system on specific hardware platforms. However, given that the machines connected to the Internet also compose the transport mechanism for the Internet, a "minority group" of server-class machines, thus affected, degraded the performance of the Net as a whole. Indeed, it can be argued that despite the greater volumes of mail generated by Melissa and LoveLetter and the tendency of some types of mail servers to achieve meltdown when faced with the consequent traffic, the Internet as a whole has proved to be somewhat more resilient in recent years.

During the 1988 mailstorm, a sufficient number of machines had been affected to impair e-mail and distribution-list mailings. Some mail was lost, either by mailers that could not handle the large volumes that "backed up," or by mail queues being dumped in an effort to disinfect systems. Most mail was substantially delayed. In some cases, mail would have been re-routed via a possibly less efficient path after a certain time. In other cases, backbone machines, affected by the problem, were simply much slower at processing mail. In still others, mail routing software would crash or be taken out of service, with a consequent delay in mail delivery. Ironically, electronic mail was the primary means that the various parties attempting to deal with the trouble were trying to use to contact each other.

In many ways, the Internet Worm is the story of data security in miniature. The Worm used "trusted" links, password cracking, security "holes" in standard programs, standard and default operations, and, of course, the power of viral replication.

"Big Iron" mainframes and other multiuser server systems are generally designed to run constantly, and execute various types of programs and procedures in the absence of operator intervention. Many hundreds of functions and processes may be running all the time, expressly designed neither to require nor report to an operator. Some such processes cooperate with each other; others run independently. In the UNIX world, such small utility programs are referred to as daemons, after the supposedly subordinate entities that take over mundane tasks and extend the power of the "wizard," or skilled operator. Many of these utility programs deal with the communications between systems. "Mail," in the network sense, covers much more than the delivery of text messages between users. Network mail between systems may deal with file transfers, the routing of information for reaching remote systems, or even upgrades and patches to system software.

When the Internet Worm was well established on a machine, it would try to infect another. On many systems, this attempt was all too easy, since computers on the Internet are meant to generate activity on each other, and some had no protection in terms of the type of access and activity allowed.

The finger program is one that allows a user to obtain information about another user. The server program, fingerd, is the daemon that listens for calls from the finger client. The version of fingerd common at the time of the Internet Worm had a minor problem: it did not check how much information it was given. It would take as much as it could hold and leave the rest to overflow. "The rest," unfortunately, could be used to start a process on the computer, and this process was used as part of the attack. This kind of buffer overflow attack continues to be very common, taking advantage of similar weaknesses in a wide range of applications and utilities.

The sendmail program is the engine of most mail-oriented processes on UNIX systems connected to the Internet. In principle, it should only allow data received from another system to be passed to a user address. However, there is a debug mode, which allows commands to be passed to the system. Some versions of UNIX were shipped with the debug mode enabled by default. Even worse, the debug mode was often enabled during installation of sendmail for testing and then never turned off.

When the worm accessed a system, it was fed with the main program from the previously infected site. Two programs were used, one for each infected platform. If neither program could work, the worm would erase itself. If the new host was suitable, the worm looked for further hosts and connections.

The program also tried to break into user accounts on the infected machine. It used standard password-cracking techniques such as simple variations on the name of the account and the user. It carried a dictionary of words likely to be used as passwords, and would also look for a dictionary on the new machine and attempt to use that as well. If an account was "cracked," the Worm would look for accounts that this user had on other computers, using standard UNIX tools.

Following the Internet Worm, and a few similar examples in late 1988 and early 1989, worm examples were very infrequent during the 1990s.

By spring 2001, a number of examples of Linux malware had been seen. Interestingly, while the Windows viruses generally followed the Christma exec style of having users run the scripts and programs, the new Linux worms were similar to the Internet/Morris/UNIX worm in that they rely primarily on bugs in automatic networking software.

The Ramen worm makes use of security vulnerabilities in default installations of Red Hat Linux 6.2 and 7.0, using specific versions of the wu-ftp, rpc.statd, and LPRng programs. The worm defaces Web servers by replacing index.html, and scans for other vulnerable systems. It does this initially by opening an ftp connection and checking the remote system's ftp banner message. If the system is vulnerable, the worm uses one of the exploitable services to create a working directory, then downloads a copy of itself from the local (attacking) system.

Lion uses a buffer overflow vulnerability in the bind program to spread. When it infects, Lion sends a copy of output from the ifconfig command,/etc/passwd, and/etc/shadow to an e-mail address in the china.com domain. Next the worm adds an entry to etc/inetd.conf and restarts inetd. This entry would allow Lion to download components from a (now closed) Web server located in China. Subsequently, Lion scans random class B subnets in much the same way as Ramen, looking for vulnerable hosts. The worm may install a rootkit onto infected systems. This backdoor disables the syslogd daemon and adds a Trojanized ssh (secure shell) daemon.

Code Red uses a known vulnerability to target Microsoft IIS (Internet Information Server) Web servers. Despite the fact that a patch for the loophole had been available for five months prior to the release of Code Red, the worm managed to infect 350,000 servers within 9 to 13 hours.

When a host gets infected it starts to scan for other hosts to infect. It probes random IP addresses but the code is flawed by always using the same seed for the random number generator. Therefore, each infected server starts probing the same addresses that have been done before. (It was this bug that allowed the establishment of such a precise count for the number of infections.)

During a certain period of time the worm only spreads, but then it initiated a DoS attack aganist http://www1.whitehouse.gov. However, since this particular machine name was only an overflow server, it was taken offline prior to the attack and no disruptions resulted.

The worm changed the front page of an infected server to display certain text and a background color of red, hence the name of the worm.

Code Red definitely became a media virus. Although it infected at least 350,000 machines within hours, it had probably almost exhausted its target population by that time. Despite this, the FBI held a press conference to warn of the worm.

Code Red seems to have spawned quite a family, each variant improving slightly on the random probing mechanism. In fact, there is considerable evidence that Nimda is a descendent of Code Red.

Nimda variants all use a number of means to spread. Like Code Red, Nimda searches random IP addresses for unpatched Microsoft IIS machines. Nimda will also alter Web pages in order to download and install itself on computers browsing an infected Web site using a known exploit in Microsoft Internet Explorer's handling of Java. Nimda will also mail itself as a file attachment and will install itself on any computer on which the file attachment is executed. Nimda is normally e-mailed in HTML format, and may install automatically when viewed using a known exploit in Microsoft Internet Explorer. Nimda will also create e-mail and news files on network shares and will install itself if these files are opened.

DETECTION TECHNIQUES

All antiviral technologies are based on the three classes outlined by Fred Cohen in his early research. The first type performs an ongoing assessment of the functions taking place in the computer, looking for operations known to be dangerous. The second checks regularly for changes in the computer system where changes should occur only infrequently. The third examines files for known code found in previous viruses.

Within these three basic types of antiviral software, implementation details vary greatly. Some systems are meant only for use on standalone systems, while others provide support for centralized operation on a network. With Internet connections being so important now, many packages can be run in conjunction with content scanning gateways or firewalls.

String Search (Signature-Based)

Scanners examine files, boot sectors, and/or memory for evidence of viral infection. They generally look for viral signatures, sections of program code known to be in specific viral programs but not in most other programs. Because of this, scanning software will generally detect only known viruses and must be updated regularly. Some scanning software has resident versions that check each file as it is run.

Scanners have generally been the most popular form of antiviral software, probably because they make a specific identification. In fact, scanners offer somewhat weak protection, since they require regular updating. Scanner identification of a virus may not always be dependable: a number of scanner products have been known to identify viruses based on common families rather than definitive signatures.

Change Detection (Integrity Checking)

Change detection software examines system and/or program files and configuration, stores the information, and compares it against the actual configuration at a later time. Most of these programs perform a checksum or cyclic redundancy check (CRC) that will detect changes to a file even if the length is unchanged. Some programs will even use sophisticated encryption techniques to generate a signature that is, if not absolutely immune to malicious attack, prohibitively expensive, in processing terms, from the point of view of a piece of malware.

Change detection software should also note the addition of completely new entities to a system. It has been noted that some programs have not done this, and allowed the addition of virus infections or malware.

Change detection software is also often referred to as integrity-checking software, but this term may be somewhat misleading. The integrity of a system may have been compromised before the establishment of the initial baseline of comparison.

A sufficiently advanced change-detection system, which takes all factors including system areas of the disk and the computer memory into account, has the best chance of detecting all current and future viral strains. However, change detection also has the highest probability of false alarms, since it will not know whether a change is viral or valid. The addition of intelligent analysis of the changes detected may assist with this failing.

Real-Time Scanning

Real-time, or on-access, scanning is not really a separate type of antivirus technology. It uses standard signature scanning, but attempts to deal with each file of object as it is accessed, or comes into the machine. Because on-access scanning can affect the performance of the machine, vendors generally try to take shortcuts in order to reduce the delay when a file is read. Therefore, real-time scanning is signifcantly less effective at identifying virus infections than a normal signature scan of all files.

Real-time scanning is one way to protect against viruses on an ongoing basis, but it should be backed up with regular full scans.

Heuristic Scanning

A recent addition to scanners is intelligent analysis of unknown code, currently referred to as heuristic scanning. It should be noted that heuristic scanning does not represent a new type of antiviral software. More closely akin to activity monitoring functions than traditional signature scanning, this looks for "suspicious" sections of code generally found in viral programs. Although it is possible for normal programs to want to "go resident," look for other program files, or even modify their own code, such activities are telltale signs that can help an informed user come to some decision about the advisability of running or installing a given new and unknown program. Heuristics, however, may generate a lot of false alarms, and may either scare novice users or give them a false sense of security after "wolf" has been cried too often.

Permanent Protection

The ultimate object, for computer users, is to find some kind of antiviral system that you can set and forget: that will take care of the problem without further work or attention on your part. Unfortunately, as previously noted, it has been proved that such protection is impossible. On a more practical level, every new advance in computer technology brings more opportunity for viruses and malicious software. As it has been said in political and social terms, so too the price of safe computing is constant vigilance.

Vaccination

In the early days of antiviral technologies, some programs attempted to add change detection to every program on the disk. Unfortunately, these packages, frequently called vaccines, sometimes ran afoul of different functions within normal program that were designed to detect accidental corruption on disk. No program has been found that can fully protect a computer system in more recent operating environments.

Some vendors have experimented with an "autoimmune" system, whereby an unknown program can be sent for assessment, and if found to be malicious, a new set of signatures is created and distributed automatically. This type of activity does show promise, but there are significant problems to be overcome.

Activity Monitoring (Behavior-Based)

An activity monitor performs a task very similar to an automated form of traditional auditing: it watches for suspicious activity. It may, for example, check for any calls to format a disk or attempts to alter or delete a program file while a program other than the operating system is in control. It may be more sophisticated, and check for any program that performs "direct" activities with hardware, without using the standard system calls.

Activity monitors represent some of the oldest examples of antiviral software, and are usually effective against more than just viruses. Generally speaking, such programs followed in the footsteps of the earlier anti-Trojan software, such as BOMBSQAD and WORMCHEK in the MS-DOS arena, which used the same "check what the program tries to do" approach. This tactic can be startlingly effective, particularly given the fact that so much malware is slavishly derivative and tends to use the same functions over and over again.

It is, however, very hard to tell the difference between a word processor updating a file and a virus infecting a file. Activity monitoring programs may be more trouble than they are worth because they can continually ask for confirmation of valid activities. The annals of computer virus research are littered with suggestions for virusproof computers and systems that basically all boil down to the same thing: if the operations that a computer can perform are restricted, viral programs can be eliminated. Unfortunately, so is most of the usefulness of the computer.

PREVENTION AND PROTECTION TECHNIQUES

In regard to protection against viruses, it is germane to mention the legal situation with regard to viruses. Note that a virus may be created in one place and still spread world-wide, so issues of legal jurisdiction may be confused. In addition, specific activity may have a bearing: in the United States: it may be legal to write a virus, but illegal to release it. However, in the first 16 years of the existence of viruses as a serious occurrence in the computing environment, only five people have been convicted in court of writing computer viruses, and in all five cases the defendants entered guilty pleas. Therefore, it is as well not to rely on criminal prosecutions as a defence against viruses.

The converse, however, is not true. If you are infected with a virus, and it can be demonstrated that your system subsequently sent a message that infected someone else, you may be legally liable. Thus, it is important to protect yourself from infection, even if the infection will not inconvenience you or cause loss to your business.

Training and some basic policies can greatly reduce the danger of infection. A few guidelines that can really help in the current environment are the following:

- Do not double-click on attachments.
- When sending attachments, be really specific when describing them.
- Do not blindly use Microsoft products as a company standard.
- Disable Windows Script Host. Disable ActiveX. Disable VBScript.
- Disable JavaScript. Do not send HTML-formatted email.
- Use more than one scanner, and scan everything.

There are now companies that will provide insurance against virus attacks. This insurance is generally an extension of "business loss of use" insurance, and potential buyers would do well to examine the policies very closely to see the requirements for making a claim against it, and also conditions that may invalidate payment.

Unfortunately, the price of safe computing is constant vigilance. Until 1995 it was felt that data files could not be used to transport a virus. Until 1998 it was felt that e-mail could not be used to automatically infect a machine. Advances in technology are providing new viruses with new means of reproduction and spread. Two online virus encyclopedias are listed in the Further Reading Section, and information about new viruses can be reliably determined at these sites.

NON-PC PLATFORM VIRUSES

As noted, many see viruses only in terms of DOS- or Windows-based programs on the Intel platform. Although there are many more PC viruses than on other platforms (primarily because there are more PCs in use than other computers), other platforms have many examples of viruses. Indeed, I pointed out earlier that the first successful viruses were probably created on the Apple II computer.

Christma exec, the Christmas Tree Virus/Worm, sometimes referred to as the Bitnet chain letter, was probably the first major malware attack across networks. It was launched on the 9th of December 1987 and spread widely on Bitnet, Earn, and IBM's internal network (VNet). It has a number of claims to a small place in history. It was written, unusually, in REXX, a scripting system used to aid with the automating of simple user processes. It was mainframe-hosted (on VM/CMS systems) rather than microcomputer-hosted, quaint though that distinction-sounds nowadays when the humblest PC can run UNIX.

Christma presented itself as a chain letter inviting the recipient to execute its code. This involvement of the user leads to the definition of the first e-mail virus, rather than a worm. When it was executed, the program drew a Christmas tree and mailed a copy of itself to everyone in the account holder's equivalent to an address book, the user files Names and Netlog. Conceptually, there is a direct line of succession from this worm to the social engineering worm/Trojan hybrids of today.

In the beginning of the existence of computer viruses actually proliferating in the wild, the Macintosh computer seemed to have as many interesting viruses as those in the DOS world. The Brandau, or "Peace" virus, became the first to infect commercially distributed software, while the nVIR virus sometimes infected as many as 30% of the computers in a given area. However, over time it has become evident that any computer can be made to spread a virus, and the fact that certain systems have more than others seems to be simply a factor of the number of computers, of a given type, in use.

CONCLUSION

Malware is a problem that is not going away. Unless systems are designed with security as an explicit business requirement, which current businesses are not supporting through their purchasing decisions, malware will be an increasingly significant problem for networked systems.

It is the nature of networks that what is a problem for a neighboring machine may well become a problem for local systems. In order to prevent this, it is critical that the information security professional help business leaders recognize the risks incurred by their decisions, and help to mitigate those risks as effectively and economically as possible. With computer viruses and similar phenomena, each system inadequately protected increases the risk to all systems to which it is connected. Each system compromised can become a system that infects others. If you are not part of the solution, in the world of malware, you are most definitely part of the problem.

GLOSSARY

Terms are derived from the "Glossary of Communications, Computer, Data, and Information Security Terms" posted online at http://victoria.tc.ca/techrev/secgloss.htm and http://sun.soci.niu.edu/~rslade/secgloss.htm.

Activity monitor A type of antiviral software that checks for signs of suspicious activity, such as attempts to rewrite program files, format disks, and so forth; some versions will generate an alert for such operations, while others will block the behavior.

Change detection Antiviral software that looks for changes in the computer system. A virus must change something, and it is assumed that program files, disk system areas, and certain areas of memory should not change. This software is very often referred to as "integrity checking" software, but it does not necessarily protect the integrity of data, nor does it always assess the reasons for a possibly valid change. Also known as authentication software when strong encryption is used.

False negative One of two types of "false" reports from antiviral software generated where no viral activity or presence is reported, even though there is a virus present; references are usually only made in technical reports. Most people simply refer to an antiviral as "missing" a virus. Also known in general security terms as a false acceptance, or Type II error.

False positive One of two types of "false" reports from antiviral software that states the activity or presence of a virus when there is, in fact, no virus. Very widely used among those who know about viral and antiviral programs, while very few use the analogous term, "false alarm." Also known in general security terms as a false rejection, or Type I error.

Heuristic Trial-and-error or seat-of-the-pants thinking rather than formal rules. In antiviral jargon: the examination of program code for functions or opcode strings known to be associated with viral activity. In most cases this is similar to activity monitoring but without actually executing the program; in other cases, code is run under some type of emulation. Recently the meaning has expanded to include generic signature scanning meant to catch a group of viruses without making definite identifications.

In the wild Viral programs that have been released into, and successfully spread in, the normal computer user community and environment; used to distinguish viral programs written and tested in a controlled research environment, without escaping, from those uncontrolled "in the wild."

Macro (virus) A small piece of programming in a simple language, used to perform a simple, repetitive function. Microsoft's Word Basic and VBA macro languages can include macros in data files, and have sufficient functionality to write complete viruses.

Malware All forms of malicious or damaging software, including viral programs, Trojans, logic bombs, and the like.

Multipartite Formerly a viral program that would infect both boot sectors and files; now a virus that will infect multiple types of objects, or that reproduces in multiple ways.

Payload The code in a viral program not concerned with reproduction or detection avoidance; often a message but is sometimes code to corrupt or erase data.

Polymorphism Techniques that use some system of changing the "form" of the virus on each infection to try and avoid detection by signature scanning software. Less sophisticated systems are referred to as self-encrypting.

Scanner A program that reads the contents of a file looking for code known to exist in specific viral programs. Also known as a signature scanner.

Stealth Various technologies used by viral programs to avoid detection on disk. The term properly refers to the technology, and not a particular virus.

Trojan horse A program that either pretends to have, or is described as having, a (beneficial) set of features but which, either instead or in addition, contains a damaging payload; most frequently shortened to Trojan.

Virus "A program which modifies other programs to contain a possibly altered version of itself" (attributed to Dr. Fred Cohen, although his actual definition is in mathematical form); "an entity which uses the resources of the host (system or computer) to reproduce itself and spread, without informed operator action."

Worm A self-reproducing program that is distinguished from a virus by copying itself without being attached to a program file, or that spreads over computer networks, particularly via e-mail; can spread without user action, for example, by taking advantage of loopholes and trap doors in software.

CROSS REFERENCES

See *Firewalls; Guidelines for a Comprehensive Security System.*

FURTHER READING

Bidgoli, H. (Ed.). (2003). *Encyclopedia of information systems.* San Diego: Academic Press.

Bontchev, V. (1994). *Are "good" viruses still a bad idea?* Retrieved April 14, 2003, from http://melona.complex.is/~bontchev/papers/goodvir.html

Cohen, F. (1994). *A short course on computer viruses* (2nd ed.). New York: Wiley.

Ferbrache, D. (1992). *A pathology of computer viruses.* London: Springer-Verlag.

F-Secure Computer Virus Information Center (n.d.). Retrieved April 14, 2003, from http://www.f-secure.com/v-descs/

Gattiker, U., Harley, D., & Slade, R. (2001). *Viruses revealed.* New York: McGraw-Hill.

Highland, H. J. (1990). *Computer virus handbook.* New York: Elsevier Advanced Technology.

Hruska, J. (1992). *Computer viruses and anti-virus warfare* (2nd ed.). London: Ellis Horwood.

IBM Research (n.d.). *Project: Antivirus research.* Retrieved January 17, 2003, from http://www.research.ibm.com/antivirus/

Kane, P. (1994). *PC security and virus protection handbook.* New York: M&T Books.

Lammer, V. (1993). *Survivor's guide to computer viruses.* Abingdon, UK: Virus Bulletin.

Shoch, J. F. & Hupp, J. A. (1982, March). The 'worm' programs—early experiences with a distributed computation. *Communications of the ACM, 25*(3), 172–180.

Slade, R. M. (1996). *Robert Slade's guide to computer viruses* (2nd ed.). New York: Springer-Verlag.

Solomon, A. (1991). *PC viruses: Detection, analysis and cure.* London: Springer-Verlag.

Solomon, A. (1995). *Dr. Solomon's Virus Encyclopedia.* Aylesbury, UK: S&S International PLC.

Sophos (n.d.). *Sophos virus analyses.* Retrieved April 14, 2003, from http://www.sophos.com/virusinfo/analyses/

Tipton, H., & Krause, M. (Eds.). (2003). *Information security management handbook: Vol. 4. Malware* (4th ed.). New York: Auerbach.

Vibert, R. S. (2000). *The enterprise anti-virus book.* Braeside, Canada: Segura Solutions.

Conducted Communications Media

Thomas L. Pigg, *Jackson State Community College*

INTRODUCTION

The Internet consists of millions of digital passages that carry signals all over the world. These conduits, which connect us to the World Wide Web, come in many shapes, sizes, and modes. The bewildering assortment and seemingly endless conduits, more commonly referred to as communication media, that are used to connect computers together can be boiled down to two types: conductive cable and wireless. Conductive media is simply a hard-wired connection, which requires someone to physically join network devices with some type of cable. The three major types of cable are coaxial, twisted pair, and fiber optic. The focus of this article will be on conducted communications media. Wireless communication is the alternative to conducted communications media. This is accomplished using a variety of broadcast transmission technologies, including terrestrial microwave and satellite communications. Wireless communications is covered in another chapter in this encyclopedia.

To aid in understanding conductive communications media, a short explanation of network transmission basics will be discussed. This will be followed by a detailed exploration of the three major types of media, mentioned above. In addition, there will be a presentation of the similarities and contrasts between these media. Finally, the suitability and application of each type of conducted media will be discussed.

OVERVIEW OF NETWORK TRANSMISSION BASICS

Network Transmission Basics

When computers communicate over the Internet or on private networks, they must follow certain communication protocols. In short, each computer that connects to a network must follow rules that govern the type of hardware and software that is used for access to the network. The protocols consist of signal types and speeds, cable layouts, and communications access schemes.

Baseband and Broadband

Bandwidth is the range of frequencies that a particular network transmission media can carry. It also reveals the maximum amount of data that can be carried on the medium. A baseband transmission is a single, fixed signal that uses the entire available bandwidth. Baseband signals use a single channel to communicate with devices on a network, which allows computers to transmit and receive data on one cable (Tittel & Johnson, 2001). Baseband communications are typically used for local area network (LANs). This environment is ideal for baseband transmissions because it is self-contained within a single location, where network traffic can be easily monitored and controlled. When LANs expand into a metropolitan area network (MAN) or a wide area network (WAN), baseband systems do not provide the bandwidths that are adequate for these larger networks. A MAN consists of several LANs within a metropolitan area. WANs are similar to MANs in that they cover a wider, possibly international, geographical area.

Network connections that attach several LANs together, such as MANs or WANs, or that allow remote access from external users or networks to local servers often require multichannel bandwidths because of increased signal traffic flow. Broadband transmissions help to meet these needs. In contrast to the discrete digital signals produced in baseband communications, broadband transmission generates an analog carrier frequency, which carries multiple digital signals or multiple channels. This concept is much like cable television systems that carry many channels on one cable.

The decision to use baseband or broadband is determined by the application. If the network consists of a single LAN, baseband would be the best choice. If an organization has computer networks that spread over several geographical areas, there might be a need for multiple channels of communications between locations, which can be achieved only through broadband applications. It is important to note that it is not necessarily the transmission media that determines whether a signal is baseband or broadband. It is the function of the devices that connect to the cable.

Cable Access Methods

A cable access method is how data is placed on the transmission media. Data is formatted into small pieces called packets. A packet contains header and trailer information about the contents, destination, error checking, and the actual data. Packets contain a specific amount of information, normally 1–2 kbytes, as defined by the particular network protocol being used. It is a lot like an envelope that is used to hold a letter to be mailed. The outside of the envelope contains a destination address along with sender information. This information is used by the United States Postal Service to deliver the mail to its destination. When a wrong destination address is used, the postal service can notify the sender by means of the return address. Network packets are handled in a similar way (Tittel & Johnson, 2001).

There are several ways in which packets are placed on the communications line. Ethernet uses a method called carrier sense multiple access with collision detection (CSMA-CD). This method is a lot like a two-way radio, which allows only one person to talk at one time. CSMA-CD requires each participating network workstation to first listen to the communication line to determine whether it is clear before attempting to broadcast a packet of information. Because of the high rates of access and speed associated with computer networks, it is very likely that two workstations may attempt to broadcast packets at the same time. If this occurs, the collision detection (CD) feature of CSMA-CD will instruct the conflicting workstations to cease transmissions and assign a random delay before they are allowed to attempt to send another packet of data. Without a random delay, the colliding workstations would continue to attempt to rebroadcast at the same time, causing more collisions. Occasionally, a defective network interface card (NIC) may cause excessive collisions by constantly broadcasting packets. This is called a broadcast storm (Tittel & Johnson, 2001).

A similar access method, used by Apple Talk networks, is called carrier sense multiple access with collision avoidance (CSMA-CA). This scheme is similar to CSMA-CD except that CSMA-CA attempts to avoid collisions altogether. Network devices accomplish this by sending a short broadcast requesting control of the network segment prior to sending a packet. This helps prevent collisions of complete packets, thus reducing the need to rebroadcast packets (Tittel & Johnson, 2001).

Another type of access method is called token passing, used by token ring and ARCNet (attached resource computer network) protocols. A token is similar to a packet. Token passing eliminates network collisions altogether by passing a token from node to node on the network in one direction. When a network device receives a token, it examines it to determine whether its contents are for that particular node. If not, the network node rebroadcasts the token to the next device on the network. When the token finds its destination, it unloads the data and rebroadcasts an empty token to the next device. The empty token will ask each passing node whether it has any data to send. If not, the empty token is passed on down the line until a device generates information to place in the empty token (Tittel & Johnson, 2001).

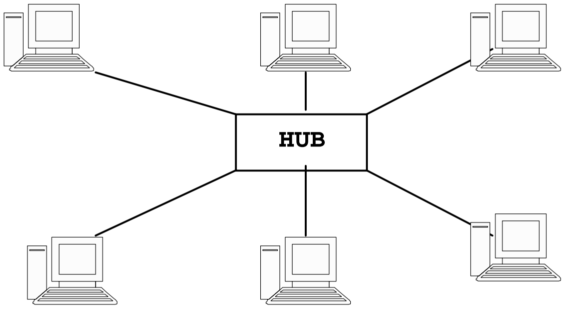


Figure 1: Star topology.

Network Topologies

A network topology is the physical cable layout or configuration. There are three basic types: star, bus, and ring. To some degree, the type of cable used will determine the type of topology used. Each network protocol (i.e., Ethernet, token ring) will specify the cable type and topology supported. The star topology, as illustrated in Figure 1, requires a central wiring point for all network devices on a particular network segment. A device called a hub or concentrator is used to serve this purpose. A hub is a multiport repeater that regenerates signals received. In some applications a device called a switch is used. A switch is more intelligent than a hub because it can learn which devices are connected to its ports and send packets directly to the destination port only, thus reducing broadcast traffic.

The bus topology (see Figure 2) requires no special external device for computers to connect to. It consists of network nodes connected to the communications media via a special T-connector that connects the computer's NIC to the bus (see Figure 3). You will notice in Figure 2 that a small box is shown at both ends of the bus. This represents a terminator that is required for this cable layout (see Figure 4). Its purpose is to prevent signal bounce, which occurs when a cable is not terminated. Signal bounce is the result of a broadcast signal that continues to oscillate. The terminator is used to absorb the signal once it has reached the end of the cable, thus eliminating signal bounce.

The ring topology (see Figure 5) is similar to bus topology in that each computer is connected directly to its neighbor without the need for a hub or other special device. The big difference between the two is that ring topology provides a closed-loop design, which does not require termination.

In some applications more than one topology may be used. For example, the token-ring protocol uses a topology called a star-wired ring, as shown in Figure 6. In this example, multistation access units (MAU or MSAU) or concentrators are used to connect the network devices. The MAUs are then connected to each other in a ring configuration.

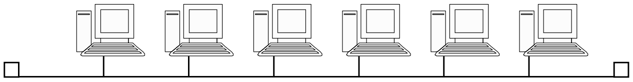


Figure 2: Bus topology.

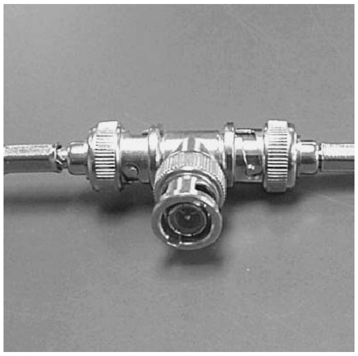


Figure 3: T-connector.

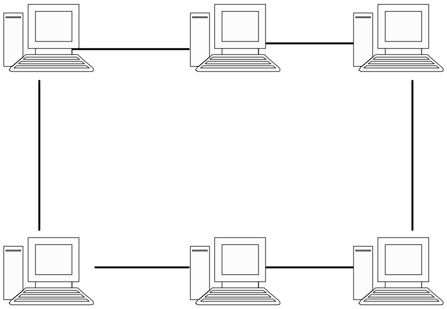


Figure 5: Ring topology.

Each topology has its advantages and disadvantages. For example, the bus topology has the advantage of low cost because it does not require any special devices to connect computers. However, a disadvantage is that if there is a break in any of the cables, the entire network segment will go down because of a loss of termination. On the other hand, computers connected to a star topology would most likely remain operating on the network if one of the cables malfunctioned between a device and a hub. Star topology requires more wire than a bus topology and it also requires hubs or switches, which is more costly. Ring topology has an advantage over bus and star tpologies because it typically provides an alternate path for data transmission in case of a cable break. Token-ring networks use this redundancy feature with star-wired ring topology.

The purpose of this section on network transmission basics is to help you understand the concepts and characteristics, discussed in the next section, related to conducted transmission cable. Each media type differs in physical and electrical/optical characteristics, as well as in application. The following section will discuss the three major types of conducted transmission cable: coaxial, twisted pair, and fiber optic.

COAXIAL CABLE

Coaxial (coax) cable is used in a wide range of applications, including television, radio, and computer network communications. This section looks at the physical makeup of coaxial cable and the specific characteristics that distinguish the different types of coax.

Components

Coaxial cable consists of an inner and outer conductor that share the same axis, which is how it got its name, coaxial (Carr, 1999). The inner conductor, sometimes called the conducting core, is typically made of a copper alloy. Most applications use a solid core, except where flexibility is required, and then a stranded core may be used. The outer conductor is either braided metal or foil, which acts as a conductor and a shield to reduce interference. An insulator, or more accurately a dielectric, is between the two conductors. Depending on the type of coax, this material is made out of Teflon or an inert gas. Finally, for protection, there is an outer sheath that surrounds the inner components. This is made from PVC (polyvinyl chloride), or special fire retardant material, for installations where wire must be run in the plenum areas above the false ceilings in a building (Carr, 1999; Tittel & Johnson, 2001).

Coaxial cables come in a variety of forms: flexible, helical line, and hard line. Flexible coax is made of flexible material in the outer conductor, typically braid or foil. This cable is mostly found in LANs and home television-to-antenna connections. Helical line coax is a semiflexible cable that consists of a slightly more rigid, spiral-wound outer conductor. Normally, this type of cable is used for network backbones or interconnections between networksthat require long cable runs. This type

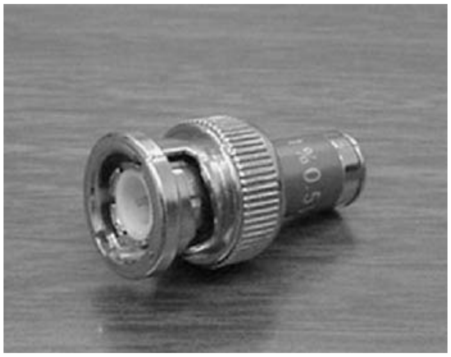


Figure 4: Terminator.

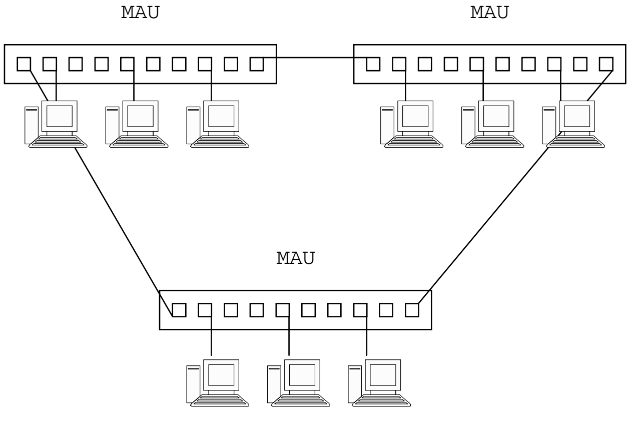


Figure 6: Star-wired ring.

of cable is able to carry a signal over longer distances at higher frequencies than can flexible cable types. Hard-line coaxial cable is used to connect equipment that transmits in the microwave frequency range. This form of coax uses a thin-walled pipe as an outer conductor, which is rigid and very difficult to work with. As the frequency of a transmitted signal increases, there is a greater chance of the signal radiating beyond the outer conductor, which also acts as a shield, thus resulting in signal loss, or in the case of external electromagnetic interference (EMI), the type of outer conductor will determine what frequencies will be rejected from the transmitted signal (Carr, 1999).

Another characteristic of coaxial cable is its impedance. Impedance is the resistance of a cable to the transmitted signal, which is measured in ohms (Tittel & Johnson, 2001). Carr (1999) states that the lowest loss of signal occurs at higher impedances; however, more power can be achieved when the impedance is low. Cable television uses coaxial cable rated at 75-ohms impedance because of its lower signal attenuation with long cable runs. Most network and radio applications use 50-ohm impedance as a middle ground between low signal loss and more power.

Cable manufacturers use a radio government (RG) specification for the coaxial cables they produce. Cable television uses either RG-6-, RG11-, or RG-59-spec cable. All of these are rated at 75 ohms. RG-6 and RG-11 are larger diameter cables used for major trunk lines. RG-59 is employed between the trunk lines and the customer's television.

RG-58 and RG-8 are 50-ohm cables used for computer network communications. As does cable television, computer networks use trunk lines to connect network segments together. RG-8 is the cable used for trunk lines or network backbones and is often referred to as thick-wire Ethernet, or thicknet for short. RG-58 is used with thin-wire Ethernet, or thinnet, to interconnect smaller segments of the network (Tittel & Johnson, 2001).

Coaxial Cable Network Applications

Thick-Wire Ethernet

Thicknet (RG-8) is typically used as an Ethernet backbone to connect various network segments together. Thicknet can be used to directly interconnect devices on the network, but usually thin-wire Ethernet is used for this purpose. Coaxial cable typically uses the bus topology. Thicknet is no exception. A thick-wire Ethernet backbone will consist of a length of RG-8 with several devices, such as hubs, connected to it throughout its length. Each device connected to the backbone requires a special device called a transceiver. The transceiver is connected to the coax by use of a vampire tap (see Figure 7). A special cable then attaches the transceiver to the device through its AUI (attachment unit interface) port (see Figure 8).

Each conducted transmission cable will exhibit certain standardized characteristics that describe its advantages and disadvantages within network applications. Table 1 describes thick-wire Ethernet characteristics. Each cable type reviewed contains similar specifications. The reason there is a maximum length is because of attenuation, which is the reduction in signal strength that occurs as cable lengths increase. At some point, the signal loss becomes excessive and causes the network to stop functioning. As with this and other wire specifications, cable manufacturers rely on IEEE (Institute of Electrical and Electronics Engineers) standards when producing their products. The Ethernet standards for thicknet and thinnet are covered in the IEEE 802.3 specification (Tittel & Johnson, 2001).

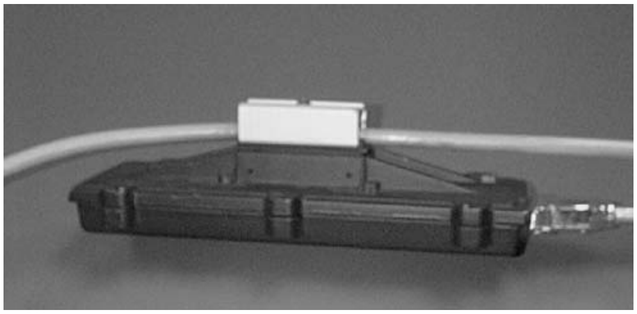


Figure 7: Thick-wire coax attached to transceiver via vampire tap.

Table 1 shows a total network cable length of 2,500 m, which is based on the 5-4-3 rule. This rule states that each network cannot exceed a total of five cable segments connected by four repeaters, with not more than three of the segments populated by nodes (see Figure 9).

Figure 9 shows five complete cable segments. Each segment is connected to a device called a repeater. A repeater is used to compensate for the attenuation caused by cable lengths that exceed 500 m for thicknet. Its function is to amplify and repeat what it receives so that its signal will be strong enough to reach another repeater or the end of the next cable segment. Note that two segments do not have any network devices connected to them. This implies that only three of the five segments may have network devices attached. Tittel and Johnson (2001) state that in reality, what this means is that any two network devices cannot be separated by more than four repeaters (five segments) with three populated segments. So, the 5-4-3 rule does not say a network is limited to only five segments. It states only that network devices cannot communicate with each other if they are separated by more than five segments.

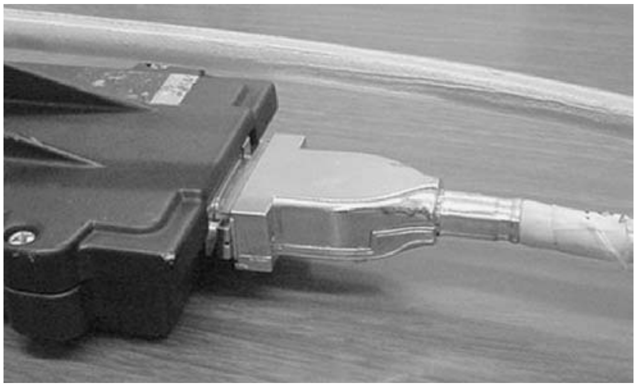


Figure 8: AUX cable connected to a transceiver.

Table 1 Thick-Wire Ethernet Specifications

Characteristic	Specification
Maximum cable segment length	500 meters
Maximum total network length	2500 meters
AUX (drop cable length)	50 meters
Maximum number of devices per segment	100
Maximum number of segments	5 connected by 4 repeaters, with not more than 3 populated segments
Bandwidth	10 Mbps (megabits per second)
Termination	50 Ohms
Cable Access Method	CSMA-CD

The AUX or drop cable length refers to the cable that attaches to the transceiver unit connected via the vampire tap to the thick-wire Ethernet backbone. This type of connection is used to connect such devices as repeaters or hubs to the Ethernet backbone. The AUX cable must be fewer than 50 m long.

Table 1 shows a maximum of 100 devices per segment for thick-wire Ethernet. Every time a device is attached to an Ethernet backbone, there will be some signal loss. Any device attached beyond the limit is not guaranteed to function properly. When designing a network, these limits should not be pushed. If a segment is fully populated, there is no room for expanding your network without adding another segment, assuming that the network is not already comprised of the maximum allowable five segments.

Some IEEE specifications, such as bandwidth, are preset and cannot be altered. If the network bandwidth will not support the needs of the network, a new protocol or standard will need to be used. The bandwidth specification of 10 Mbps for thick-wire Ethernet describes how much data can be transmitted, which ultimately determines the speed of communications. As with all bus topologies using coax, a terminator is required at both ends of the bus to eliminate signal bounce. Because thicknet coaxial cable is designed for 50-ohm impedance, a 50-ohm terminator is required. A terminator is simply a connector that has been retrofitted with a 50-ohm resistor between the inner and outer conductors.

Finally, Ethernet packets are transmitted using the CSMA-CD cable access method. You will recall that this manner of access is a lot like a two-way radio, which requires the user to monitor the radio for traffic before he or she attempts to talk or transmit. This is also true of CSMA-CD, except it is the responsibility of the NIC and the communications software to monitor the network traffic.

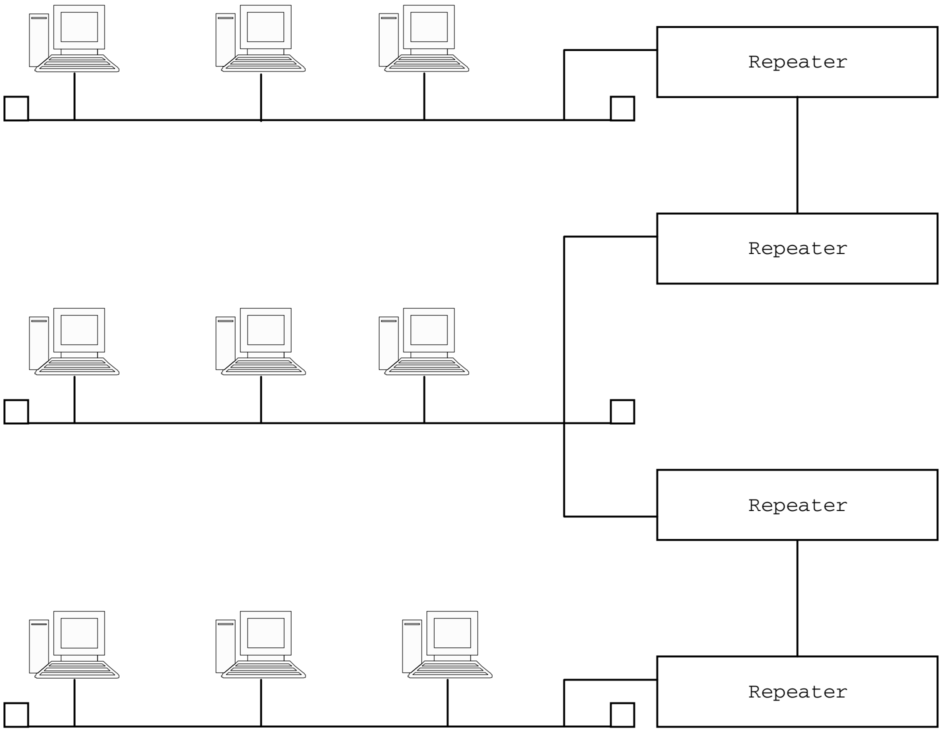


Figure 9: 5-4-3 rule diagram.

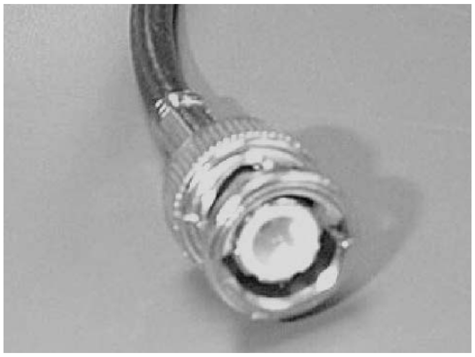


Figure 10: BNC connector.

The collision detection (CD) feature provides a means of rebroadcasting a packet of data when the occasional collision occurs.

Thinwire Ethernet

A network can be quickly set up using very little more than RG-58 cable and a few connectors. Thin-wire cable is commonly used to interconnect computers or other network devices on a network segment using the bus topology. In the past, thinnet was one of the most popular transmission cables for Ethernet networks because of its low cost and ease of set up.

Thinnet cable uses a special connector, called a BNC (British naval connector or bayonet nut connector) connector (see Figure 10). Computers and other network devices are connected to the bus topology by way of a BNC T-connector (see Figure 3), which connects to the NIC. The NIC is a transceiver that provides the physical connection between the network device and the transmission media. In addition, the NIC is responsible for packaging the data coming from the network device into a form acceptable to the network cable.

As with thicknet, thinnet follows the IEEE 802.3 standard with regard to the use of RG-58 coaxial cable. Table 2 summarizes these cable specifications. You will notice that thinnet characteristics are similar to thicknet except in cable lengths and number of devices per segment. Also, there is no specification for a drop cable for thin-wire Ethernet because the NIC card serves the same function as the external transceiver used by thicknet (Tittel & Johnson, 2001).

The physical characteristics of thin-wire coax attenuates signals over shorter distances than does thick wire. This is not really a big problem because thinnet is normally used to interconnect computers within a relatively small area. In situations where computers need to be separated by more than 185 m, a repeater can be used to accommodate these devices. The low cost of thin wire far outweighs the distance limitations.

The number of devices that can be attached to a thinnet bus topology is 30. This is fewer than can be attached to thicknet because of the reasons noted in the previous paragraph for cable length. Be aware that a network can have more than one segment. With repeaters, thin-wire Ethernet can support multiple populated segments. The 802.3 standard specifies a maximum of 1,024 devices per network as long as the 5-4-3 rule is followed (Tittel & Johnson, 2001).

In addition to network protocol standards for Ethernet and token ring, IEEE references cable-type standards such as 10Base5 and 10Base2 for thick wire and thin wire, respectively. The number 10 represents the bandwidth of the cable specification. Base refers to a baseband (single-channel) signal type and the last number represents the maximum length of a network segment in meters rounded to the nearest hundred. For example, 10Base5 refers to a coaxial cable that carries a baseband signal transmitted at 10 Mbps with a maximum segment length of 500 m (thicknet). Thinnet is referenced by 10Base2, meaning a baseband signal with a 10-Mbps bandwidth at 200 m, approximately. In reality, thin-wire Ethernet can only support 185-m segments. Twisted-pair and fiber-optic cables carry 10BaseT and 10BaseF IEEE cable standards. These are discussed in more detail later.

One other coaxial cable application that has decreased in use over the past few years is ARCnet. ARCnet uses an RG-62, 93-ohm impedance, coaxial cable. This was a fairly popular media back in the 1980s, but because of its low bandwidth (2.5 Mbps), it is not used as much anymore. One of its major benefits was its support of longer cable segments, 610 m (Tittel & Johnson, 2001).

There are several advantages to using coaxial cables for computer networks. For LANs, RG-58 uses the bus topology, which offers lower cost options and ease of installation. Also, the outer shielding of coax provides moderate protection from interference. 10Base2 applications allow for fairly long cable runs compared with other cables, such as twisted pair. The major drawback of coax is its narrow bandwidth compared with contemporary applications using twisted-pair and fiber-optic cable. Plus, thicknet installations can be somewhat difficult and

Table 2 Thin-Wire Ethernet Specifications

Characteristic	Specification
Maximum cable segment length	185 meters
Maximum total network length	925 meters
Maximum number of devices per segment	30
Maximum number of segments	5 connected by 4 repeaters, with 3 populated segments
Bandwidth	10 Mbps (megabits per second)
Termination	50 Ohms
Cable Access Method	CSMA-CD

expensive because of the rigidity of the cable, the rather difficult task of installing network connections, and the overall cost of the interface devices.

TWISTED-PAIR CABLE

Twisted-pair cable consists of one or more pairs of twisted wire. Twisted pair has a longer history than coaxial cable, but its usage was restricted to voice only until the 1980s. Omninet or 10Net began to use twisted pair in the early 1980s for PC-based LANs. One of the first PC applications using twisted-pair cable was in 1984, when IBM introduced the token ring network protocol. By the end of the 1980s, Ethernet technology began using twisted pair. Some advantages of twisted-pair over coax cable include its lightweight and flexibility, and in some cases, existing twisted-pair cables within buildings can be used without having to install new wiring (*Network Magazine,* 1999).

Components

Twisted-pair media comes in one of two forms: STP (shielded twisted pair) or UTP (unshielded twisted pair). STP consists of several pairs of twisted wires that are surrounded by a foil shielding and an outer jacket or sheath. The number of wire pairs can vary from either two or four for basic telephone and network applications, to hundreds of pairs for major communication trunk lines. Each wire within the pair consists of a solid or stranded copper center core surrounded by an insulating cover. Stranded wire is typically used where greater flexibility is required.

The reason for the twist in each pair of wire is to reduce cross talk. When a signal travels down the wire it produces a magnetic field. If not controlled in some way, this can produce unwanted interference with other pairs of wire within the same cable (cross talk). The twist in the pairs works to reduce the effects of cross talk. It also helps to reduce the magnetic field's effect on other pairs. Cables with more twists per foot offer the best performance and are normally more expensive. The foil shielding that surrounds the twisted pairs block sources of interference from electronic noise outside of the cable in addition to keeping stray noise from escaping to the outside world. STP is often the choice of transmission media when heavy electrical or electronic equipment exists near cable runs.

The outer sheath exists to protect the inner components of the cable. Like many cables, twisted-pair jackets are made of some form of PVC or special plenum sheaths that are used in applications that require nontoxic and fire-resistant cable.

UTP is basically the same as STP except that it does not have a foil shield (see Figure 11). Without the shield, it is more prone to cross talk and other forms of interference, although some cross talk is reduced because of the twists in the pairs of wire.

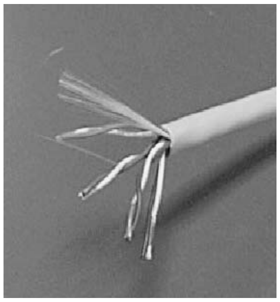


Figure 11: UTP cable.

UTP is probably the most popular cable type for networks today. The Electronic Industries Alliance (EIA) and the Telecommunications Industries Association (TIA), with the endorsement of the American National Standards Institute (ANSI) have defined several categories of UTP cable for different applications (Tittel & Johnson, 2001; *Network Magazine,* 1999; Fogle, 1995). Category 1 is traditionally used for telephone/voice transmissions. Category 2 is capable of data transmissions of up to 4 Mbps. Category 3 is certified for up to 10 Mbps for 10BaseT Ethernet and suitable for 4-Mbps token ring. New technology applications, including 100BaseT4 and 100Base-VG AnyLAN, can utilize this lower-rated cable for 100-Mbps bandwidth. Category 4 can handle up to 16 Mbps, which includes token ring 16 Mbps and 10-Mbps Ethernet applications. Category 5 cable type supports bandwidths of up to 100 Mbps. This is probably the most popular cable in use today. Category 5E cable has been tested up to 400 MHz and supports 1000 Mbps under the 1000BaseT (gigabit Ethernet) standard. Category 6, 6E, and 7 are newly developed or in the development stage. These cables will basically support higher bandwidths, further reductions in cross talk, and increased stability at frequencies above 500 MHz (Global Technologies, 2002).

Twisted-Pair Cable Network Applications

Twisted-pair cable uses a special modular connector. Standard telephone wire uses a four-connection plug called an RJ-11. For network applications, an eight-connector RJ-45 is used (see Figure 12). UTP cables often consist of a solid core with a thin layer of insulation. RJ-45 connectors, called insulation displacement connectors (IDCs), require a special tool that pushes the insulated wire into the connector. When each wire is pushed into the connector, the insulation is sliced just enough to make contact with the center core of the wire (Spurgeon, 2000).

Twisted-pair cable is typically used for point-to-point wiring, such as with star topology configurations. UTP is connected to a network device via a NIC card's RJ-45 jack (see Figure 13), with the other end connected to a hub, switch, or router. In some cases a patch panel or

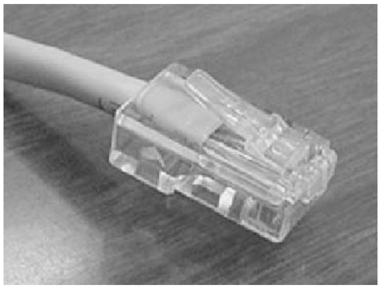


Figure 12: RJ-45 connector.

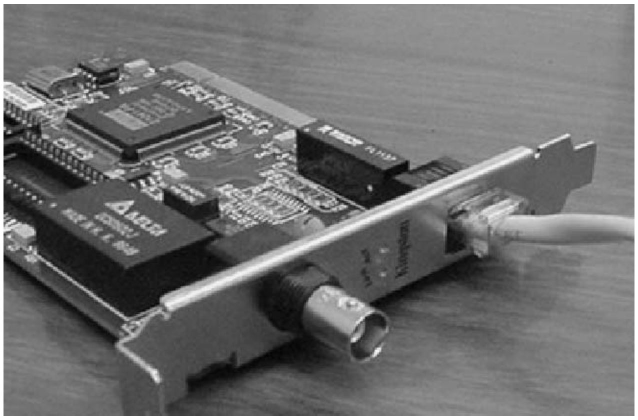


Figure 13: RJ-45 connection to a NIC.

punch-down block may be used to better organize cabling before it is attached to the devices.

There are several IEEE Ethernet standards for twisted-pair media. The most documented are the 10BaseT, 100BaseT, and 1000BaseT standards (see Table 3). All twisted-pair Ethernet standards require a maximum cable length of 100 m and use star topology, which allows only one network connection for each cable segment. You may note that this is considerably less than the coaxial cable specifications. It is much more difficult to control high-frequency signals over twisted pair; thus, shorter cable runs are used to cut down on the effects of a harsher electrical environment (Spurgeon, 2000). The major differences between 10BaseT, 100BaseT, and 1000BaseT are their bandwidths: 10, 100, and 1000 Mbps, respectively.

Each IEEE cable type specifies which wire type is most suitable for a particular application. Category 3 cable is adequate for 10BaseT applications. In fact, only two of the four pairs of wire are actually used. The 100BaseT standard is a bit more complicated. There are three cable configurations for the 100BaseT standard: 100BaseT4, 100BaseTX, and 100BaseFX. 100BaseT4 can produce 100-Mbps bandwidth using all four pairs of category 3, 4, and 5 cables. With 100BaseTX, only two pairs of category 5 cable are used to transmit and receive at 100 Mbps. 100BaseFX is the fiber optic extension for 100BaseT. 100BaseT is often referred to as fast Ethernet. One of the newest technologies, gigabit Ethernet, transmits 250 Mbps over each of the four pairs of category 5 wire simultaneously for a total bandwidth of 1 Gbps (Tittel & Johnson, 2001).

Ethernet standards for UTP follow TIA/EIA 568 standards. Wire pairs within a particular category of cable are color coded. The preferred 568 standard dictates that wire pairs will consist of a green wire paired with a white wire with a green stripe, an orange wire paired with a white wire with an orange stripe, a blue wire paired with a white wire with a blue stripe, and a brown wire paired with a white wire with a brown stripe. Some cable manufacturers will put a white stripe on top of the solid color within each pair. Each pair will be attached to a specific connection on the RJ-45 connector (see Figure 14). Tip and ring are used to identify the connections made by each pair. These terms come from the old telephone systems, which required a patch cable to connect one telephone line to another. The connector on the patch cable consisted of a tip at the end and a ring separated by an insulator that separated the two conductors. These two conductors provided the path for transmitting or receiving communications. Traditional analog telephone lines require only one pair for operation. Network transmissions typically require two or four pairs. Note the discussion above regarding the required pairs for each of the Ethernet standards 10BaseT, 100BaseT, and 1000BaseT (Spurgeon, 2000).

The majority of network device connections using UTP use straight-through wiring, which means that both ends of the RJ-45 connector are wired exactly the same way. Each pair of cable either transmits or receives data. Likewise, each NIC card transmits data through certain connections and receives through others via its RJ-45 port. If a straight cable was connected to two NIC ports, the result would be that the transmit connections of both ports would be connected to each other, as would the receive links. It should be apparent that when one NIC card transmitted, the other would not receive it because of the direct connections between the ports. Therefore, there must be some kind of crossover that makes a connection between the transmit and the receive pairs. In reality, most network connections between devices are made via hubs or switches. These devices provide the crossover for the transmit and receive pairs.

There are some applications where a special cable called a crossover cable is required. A crossover cable is made by crossing pairs two and three on the opposite end

Table 3 Ethernet Twisted-Pair Cable Specifications

Characteristic	Specification
Maximum cable segment length	100 meters
Maximum number of devices per segment	1
Bandwidth	
10BaseT	10 Mbps
100BaseT4, 100BaseTX	100 Mbps
1000BaseT	1000 Mbps
Topology	Star
Cable Access Method	CSMA-CD

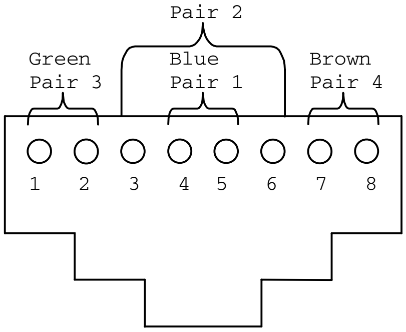


Figure 14: T568 wiring scheme.

of the connecting cable. This allows you to connect two network devices together independently without using a hub or other central wiring point. For example, if you had a small office with two computers that you wanted to network, you could build or buy a crossover cable and connect the NICs from both computers together to form a small network without the expense of a hub. Crossover cables are also used to connect hubs or switches together if they are not supplied with an uplink port. Uplink ports are optional ports that take care of uncrossing the links between these centralized wiring points.

In comparison, there are several advantages and disadvantages to twisted-pair and coaxial cabling. By far, coax has the distance contest won; however, it does not seem that coax has been keeping up with the network cabling race with regard to bandwidth. During the mid-to-late 1980s, coax was the wire of choice, but the torch has now been passed to twisted pair in the LAN interconnection race. Twisted-pair cable offers great flexibility within network design and layout. Even though maximum cable runs are shorter than coax, 100 m is still a significant distance for most network connections. With applications being developed to use the newest categories of UTP, categories 6 and 7, the potential advances for twisted-pair cable are very encouraging with regard to increased bandwidth and expanded connection options.

Token Ring Cabling

In the mid-1980s, IBM developed the token ring architecture. This network protocol used a special type of twisted-pair wiring. IBM engineered a cabling system that included nine different cable types, numbered from one to nine. These cable types included STP, UTP, and fiber optic mediums. For token ring networks, types 1 (STP) and 3 (UTP) were the most prevalent. Type 1 cable allowed for lengths of up to 101 m, with a maximum bandwidth of 16 Mbps. Because of its lack of shielding, type 3 cable runs were limited to 45 m and a bandwidth of 4 Mbps (Tittel & Johnson, 2001). Today, most token ring applications use UTP category 3 or 5 cable with RJ-45 connectors. As noted in Table 4, there have been some improvements in cable lengths with the use of this media (Feit, 2000).

The IEEE specification for token ring is 802.5. This standard describes a bandwidth of 4 or 16 Mbps using the token passing cable access method, which is described in the first part of this chapter. Token ring uses a star-wired ring topology (see Figure 6). Fault tolerance is probably the biggest advantage the token ring architecture has to offer. The star-wired ring topology, using the IBM cable system, provides a dual path for data transmission. Theoretically, if a cable is cut, the token ring network can continue to operate because of the redundant path. Another plus is the use of the token passing access method, which all but eliminates collisions (Tittel & Johnson, 2001; Feit, 2000).

Table 4 Token Ring Cable Specifications

Characteristic	Specification
Maximum cable segment length	101 meters (type 1) 45 meters (type 3) 100 meters (Cat 3) 225 meters (Cat 5) 400 meters (STP)
IEEE specification	802.5
Bandwidth	4/16 Mbps
Topology	Star-wired ring
Cable access method	Token passing

FIBER OPTIC CABLE

Consider that the capacity of a simple telephone line to carry data is equivalent to the flow of liquid through a drinking straw. It would take a tunnel large enough to drive a bus through to equal the bandwidth of a single fiber optic strand using today's technologies. Furthermore, if technology existed to fully utilize fiber optic capabilities, it would take a tunnel the diameter of the moon to handle the flow of liquid equal to its potential bandwidth (Jamison, 2001).

Fiber optic cable carries optical pulses rather than electrical signals. An optical signal virtually eliminates any possible electrical interference. The potential for electronic eavesdropping is completely eradicated. Fiber optic cable lengths are measured in kilometers rather than meters, which greatly reduces the number of signal repeating devices required for long-distance cable runs. One of the most impressive characteristics is the high volumes of data that can be pumped through optical fibers. All this makes fiber optic cable the medium of choice for many WAN applications, where network traffic can cause major bottlenecks (Tittel & Johnson, 2001). In addition, there are many people who believe that fiber optic cables may become the norm rather than the exception for interconnecting network devices on a LAN (Kostal, 2001; Sinks & Balch, 2001; Vickers, 2001).

Components

Fiber optic cable consists of three components: a fiber core, cladding, and a protective outer sheath. The inner core is either glass or plastic fiber. Plastic is used when a flexible cable is needed; however, plastic cable is more susceptible to attenuation than glass, which limits the length of the cable. Cladding provides a coating the keeps the light waves within the fiber. The outer protective jacket is used to protect the fiber inside (Tittel & Johnson, 2001).

Fiber Modes

Fiber optic cable comes in two forms, single and multiple mode. Single-mode fiber is best for applications that require great distances. This fiber is very narrow and requires an extremely focused source of data transmission that can keep the light waves on a straight, continuous path. The narrow internal core makes room for one-way traffic only. This reduces obstacles (stray light) that might interfere with the delivery of packets, resulting in more reliable transmissions over longer distances. A laser is normally used as the generating source for single-mode transmissions because it can produce the narrow band of light needed (Jamison, 2001).

Multimode fiber (MMF) is a cheaper alternative to single-mode because of lower cost transmission devices such as LEDs (light emitting diodes). MMF has a larger core than does single mode, which allows more room for the light signal to travel, allowing lower cost transmission devices. With a wider core and less precise light-emitting sources, there is more information loss due to light dispersion. A larger core allows the light wave to bounce around, which results in shorter transmission distances. However, the connections to the fiber are easier to make to the larger core. MMF is ideal for shorter cable runs within single buildings for LAN backbones (Jamison, 2001).

Fiber Optic Cable Network Applications

There are four major types of fiber connectors: straight (ST), straight connection (SC), medium interface connector (MIC), and subminiature type A (SMA). ST is often used for interconnections between individual fibers and optical devices. When joining optical fibers, an SC connector may be used. This connector consists of connections for two fibers: one for receiving and the other for sending. MIC connectors are one-piece connectors similar to the SC used to connect both the transmit and the receive fibers. It is primarily used with the fiber distributed data interface (FDDI) protocol. As with the ST connector, SMA uses individual connectors for each fiber. The major difference is that SMA uses either a straight or a stepped ferrule in order to ensure a precise fit, whereas the ST connector uses a bayonet twist-lock connection (Tittel & Johnson, 2001).

The major differences between fiber optic and other cable types are fiber optics long-distance cable runs and high bandwidth. For example, the maximum fiber length for the 100BaseFX Ethernet standard is 2000 m and the bandwidth for 1000baseSX tops out at 2000 Mbps. These specifications easily exceed other cable specifications. The potential of fiber optic cable is limited to the technology that exists today.

There are applications where fiber optic is needed not for its high bandwidth but to facilitate long cable lengths or to block unwanted electrical interference. For example, the 100BaseFX specification supports the use of fiber optic cable. With the exception of distance, the other specifications for 100BaseFX are essentially the same as that described for UTP cabling (see Table 5). Token ring also has a cable specification for the use of fiber optic for connecting its MAUs (MSAUs). Again, this does not fully utilize fiber optic's full bandwidth potential, but it allows for longer cable runs.

Table 5 Fiber Optic Cable Specifications

Characteristic	Specification
Maximum cable segment length	2000 meters
IEEE specification	802.3
Bandwidth	100 Mbps
Topology	Star
Cable access method	CSMA/CD

Fiber optic applications do not stop at 100 Mbps. Gigabit Ethernet, 1000BaseLX and 1000BaseSX, each sport bandwidths of 1000 Mbps. The major difference in these two standards is the type of laser used. 1000BaseLX uses a long-wavelength laser that can transmit a signal over 5000 m of fiber in full-duplex mode. A short-wavelength laser is used in 1000BaseSX applications, which support a maximum cable length of 550 m and 2000 Mbps bandwidth in full-duplex mode (Tittel & Johnson, 2001).

The only thing holding fiber optic transmission rates at bay is the devices used to drive the signals through the fibers. As noted at the beginning of this section, the potential of fiber optic media is truly grand. Currently the primary application for fiber optic cabling is for long connections and network backbones. There are some limited applications of direct connections between individual computers and hubs, but many people feel that this will increase over time as cost continues to decrease and installation of fiber becomes easier.

COMPARISONS AND CONTRASTS

When comparing and contrasting the advantages and disadvantages of the different types of cables, you must look at the particular application. It may be true that fiber optic cable has the potential to transmit data much faster than UTP; however, if bandwidth is not an issue, the added complexity involved in installing fiber optic where it is not needed would make it a bad choice. One must look at all the characteristics of the transmission medium before choices are made. In some cases, a network may require a mixture of cables to meet specific needs for data transmission. Table 6 compares and contrasts each cable type discussed in this article, showing more clearly how a particular cable can applied (Tittel & Johnson, 2001).

CONCLUSION

Conductive transmission cables provide the highway for data communications. This article has discussed the most

Table 6 Comparison of Cable Types

Cable	Bandwidth	Length	Interference	Installation	Cost
Thinwire coax	10 Mbps	185 m	Moderate	Easy	Low
Thickwire coax	10 Mbps	500 m	Low	Hard	High
UTP	100 Mbps	100 m	High	Easy	Low
STP	1000 Mbps	100 m	Moderate	Moderate	Moderate
Fiber optic	10 Gbps	100 km	None	Rather hard	Expensive

popular types of cable in use today. As times change, the applications and uses for these different cable types will change. The most limiting factors for conductive transmission media are the devices that drive the packets of information through them. We saw that bandwidths for twisted pairs have increased from around 1 to 1000 Mbps. Transmissions speeds over fiber optic cable are only now being realized. What are the limits? What does the future hold for conductive transmission cables? How do wireless technologies fit into network communications? No one knows the answers to these questions, but at least the questions are being asked, so researchers can continue to experiment and produce new applications for all transmission options.

GLOSSARY

Bandwidth The maximum range of frequencies a communications media can carry.

Cross Talk An electromagnetic field surrounding certain types of cable that may interfere with adjacent wires.

Hub A central wiring point for network devices configured in a star topology. Its purpose is to repeat the broadcast packets to all network nodes.

Impedance The resistance of conducted communications media to a transmitted signal.

Insertion Loss The loss or attenuation of a signal that occurs each time a device is inserted into a network using conductive communications media.

LAN Local area network; a collection of computers connected to a network within a single building or location.

NIC Network interface card; an electronic device that is installed in a network node that provides a link to the network media.

Node Any device connected directly to a network.

Packet Contained data, destination, origination, and error-checking information that is transmitted over a network.

Protocols A set of rules or specifications for different types of network communications.

Signal Attenuation The loss of signal caused by increasing lengths of cable.

Terminator A device attached to a cable configured with a bus topology that eliminates signal bounce.

Topology The physical cabling configuration of a network.

CROSS REFERENCES

See *Circuit, Message, and Packet Switching; Local Area Networks; Propagation Characteristics of Wireless Channels; Public Networks; Radio Frequency and Wireless Communications; Standards and Protocols in Data Communications; TCP/IP Suite; Wide Area and Metropolitan Area Networks*.

REFERENCES

Cabling (1999). *Network magazine*. Retrieved April 2, 2002, from http://www.networkmagazine.com/article/NMG20000724s0010

Carr, J. J. (1999). Coax ‘n’ stuff. *Popular Electronics, 16*(9), 77–79.

Feit, S. (2000). *Local area high speed networks*. Upper Saddle River, NJ: New Riders Publishing.

Fogle, D. (1995). *Network magazine*. Retrieved April 2, 2002, from http://www.networkmagazine.com/article/NMG20000724s0011

Global Technologies, Inc. (2002). Retrieved April 15, 2002, from http://www.globaltec.com/catext100.html

Jamison, E. (2001). Finding out about fiber. *Poptronics, 2*(11), 21–23.

Kostal, H. (2001). Switching to all-optical networks. *Lightwave, 18*(13), 106–108.

Sinks, C., & Balch, J. (2001). Fiber snakes its way closer to the desk. *Lightwave, 18*(11), 86–88.

Spurgeon, C. E. (2000). *Ethernet: The definitive guide*. Cambridge, MA: O’Reilly.

Tittel, E., & Johnson, D. (2001). *Guide to networking essentials*. Toronto, Ontario, Canada: Course Technology.

Vickers, L. (2001). Emerging technology: Is fiber optic destined for the desktop. In *Network magazine*. Retrieved April 2, 2002, from http://www.networkmagazine.com/article/NMG20010103S0004

Consumer Behavior

Mary Finley Wolfinbarger, *California State University, Long Beach*
Mary C. Gilly, *University of California, Irvine*

INTRODUCTION

Despite the recent shakeout in Internet retailing, particularly pureplays, online shopping continues to grow, from $42.4 billion in 2000 to $47.6 billion in 2001 (Forrester Research, 2002); by 2005, business-to-consumer (B2C) revenue is forecasted to be $156 billion. In May 2002, the number of monthly active Internet users in the United States was close to 107 million (Nielsen-Netratings, May 2002), and Jupiter Media Metrix estimates that the number of those users who are online shoppers is currently 67 million, a number predicted to reach 132 million by the year 2006. Moreover, the demographics of online shoppers are increasingly expected to reflect offline demographics; as well, shoppers will eventually make higher priced expenditures and buy items, like apparel, that in the past have been considered too risky to purchase online. Jupiter Media Metrix also predicts that over the next 5 years, the majority of online shoppers in the United States are more likely to be older than 35 and to have incomes of $30,000 to $75,000, compared with the current online buying population, which is somewhat younger and more affluent (Tedeschi, 2002).

Although the United States is the world's largest market for online shopping, advanced European markets such as Germany and Britain are expected to grow as well (Center for Media Research, 2001). Forrester Research predicts that the online retail sector in Europe will be 32.8 billion Euros in 2002; moreover, European consumers are expected to increase their online spending more than 50% in 2003 compared with 2002 (Pastore, 2002). According to the annual *World E-Commerce and Internet Market Report* (WECIM), produced by the META Group, northern European countries in particular have adopted technology faster than southern European countries (Pastore, 2001a). In Asia, South Korea and Malaysia are emerging as leaders in the area of e-commerce. Other countries with a growing online e-commerce presence include Canada, Australia, and New Zealand.

So, e-commerce continues to increase in the United States and is spreading to many countries around the world. But what do we know about the wants, needs, and experiences of online consumers? In this chapter, we address the factors that predict online shopping and discuss motivations for online shopping. Next, we present Technographics, a segmentation scheme that classifies U.S. consumers based on their attitudes toward technology, their primary use of the Internet, and their income. As well, we review what is known about the attributes that consumers desire when using websites to make online purchases. We conclude by speculating about the future of online shopping.

FACTORS PREDICTING ONLINE SHOPPING

Scholars and market researchers have suggested several factors that make shopping online more likely for a consumer. Compatibility with a consumer's lifestyle plays a strong role, with positive attitudes toward technology, adoption of multiple kinds of new technologies, online skill, and longer online experience all predicting a stronger likelihood of living a wired lifestyle, including making online purchases (Bellman, Lohse, & Johnson, 1999; Lohse, Bellman, & Johnson, 2000; Modahl, 2000; Hoffman, Novak, & Schlosser, 2000). Previous use of catalogs also predicts that online shopping will be more compatible with a consumer's lifestyle; consistent with this line of reasoning, consumers utilizing catalogs were among the early adopters of online shopping (Pastore, 1999). Those consumers who feel that they must physically examine products are less likely to shop online (Li, Kuo, & Russell, 1999). Over time, visual experiences have been consistently improved online so that products can be better examined; nevertheless, haptic devices, which might give consumers the experience of touching objects, are only in the very earliest stages of development.

Consumers with goal-oriented personalities, or in other words those with an internal rather than an external locus of control, are more likely to shop online (Novak, Hoffman, & Yung, 2000). Perhaps goal-oriented consumers feel more time starved; research suggests that time starvation also predicts the likelihood of adopting online shopping. Partially because higher income, career-oriented consumers are more likely to feel that they are time starved, they are more likely to shop online (Modahl, 2000). Situational factors also enter the picture, including type of product; computers and software, travel, music and videos, and books have the highest category penetration rates (Silverstein, Stanger, & Abdelmessih, 2001). Shopping mind-set also affects the propensity to shop online, as consumers are more likely to choose online shopping when their motivations are goal-directed rather than experiential (Wolfinbarger & Gilly, 2001). Consistent with the association of motivations with shopping outlet, toy retailer KB Toys has found that more impulsive experiential shoppers purchased toys at their land-based stores whereas goal-focused consumers were more likely to shop online; thus, the shopping outlets compliment rather than compete with each other. Following, we discuss in more detail how goal-oriented and experiential motivations affect the tendency to shop online.

CONSUMER MIND-SETS AND ONLINE SHOPPING: GOAL-ORIENTED VERSUS EXPERIENTIAL SHOPPING ONLINE

Consumer shopping mind-sets may be experiential or goal-oriented (Babin, Darden, & Griffen, 1994); both mind-sets are found in online as well as offline shopping. However, online buyers are more likely to be goal focused rather than experiential; nevertheless, some online shoppers are experiential and are thus shopping for fun. Experiential behavior is more likely in categories where shoppers have an ongoing, hobby-type interest, where the search for that special something is as enjoyable as adding to the collection. As well, when consumers feel they have time available to surf to see what's "out there," the result is more experiential shopping behavior. Research indicates that the enjoyment associated with experiential behavior results in a number of benefits for online marketers. These benefits include a more positive mood, greater shopping satisfaction, and a higher likelihood of impulse purchasing.

Goal-oriented or utilitarian shopping is efficient and deliberate, with a preplanned purchase in mind. Thus, goal focused shoppers "buy" rather than "shop" and want to do so quickly and without distraction. Two thirds to four fifths of Internet buyers are goal-oriented, performing narrowly defined searches for specific products online rather than browsing (Solomon, 1999; Wolfinbarger & Gilly, 2001). Analysis of clickstream data of major e-commerce sites also suggests that most online consumers are goal-focused, as goal-oriented shoppers are expected to spend less time at a site (Hoffman & Novak, 1996). In topline clickstream data posted weekly at Nielsen-NetRatings (http://www.nielsen-netratings.com), the average length of visits at major e-commerce sites (with the significant exception of e-Bay) is shown largely to be 10 min or less.

E-tailing consumers are expected to be goal focused for several reasons. First, time-starved consumers are especially likely to be online shoppers (Bellman, Lohse, & Johnson, 1999), going online to buy rather than browse. Second, early and heavy users of the Internet tended to have a strong internal locus of control, resulting in goal-oriented personalities who express these personalities in their online shopping behavior (Hoffman, Novak, & Schlosser, 2000). Also, search costs for product information are dramatically reduced online, facilitating goal-oriented searches (Klein, 1998).

Research of online shopping behavior suggests that goal-oriented shoppers are interested in e-tailing because it offers the following: (a) convenience and accessibility, (b) selection, (c) availability of information, and (d) lack of sociality. Further, shoppers associate these goal- oriented attributes explicitly with increased freedom and control over their purchase transactions (Wolfinbarger & Gilly, 2001).

Although goal-oriented motivations predominate in online shopping, some online shoppers do engage in experiential shopping, or shopping for fun. These experiential shoppers enjoy (a) auctions, (b) involvement with a hobby/product class, and (c) bargain hunting; thus, these experiential shoppers usually focus on the enjoyment of shopping as much or more than they focus on product acquisition (see Tables 1–4).

Goal-Oriented Shopping Behavior

Goal-oriented shoppers frequently talk about the freedom and control they experience while shopping online and how much more freedom and control they have online relative to offline. Consumers often decide to go online to shop only when seeking a specific purchase, with most describing online buying as consisting almost entirely of planned purchases. In contrast, shoppers overwhelmingly report that they tend to be more impulsive offline than online (except at auction sites, with their experiential attraction). The tendency to be rational rather than impulsive during online shopping arises from several factors, including the inability to obtain goods immediately, the ease of returning virtually whenever they want to buy the goods after further thought (no driving or walking), and the inconvenience of having to return unwanted items. The ease of getting online to buy whenever they decide they really want an item appeals to consumers, decreasing their sense of commitment to making a purchase during a particular visit and increasing their sense of control over purchase situations (Wolfinbarger & Gilly, 2001).

In summary, goal-oriented shoppers appreciate the freedom and control coupled with the lack of commitment to purchase immediately that they experience in the online environment. They typically experience little prepurchase pressure and are thus less likely to be impulsive. It is important that goal-oriented consumers particularly associate four attributes of online shopping—convenience,

Table 1 Motivations for Shopping Online

Shopping Mood	Important Factors	Outcome Desired
Goal-oriented	Accessibility/convenience Selection Information availability Lack of sociality	Freedom, control Commitment to goal, not experience
Experiential	Involvement with product class Positive sociality Positive surprise Bargain hunting	Fun Commitment to experience as important or more important than goal

Note. Source: Wolfinbarger & Gilly (2001).

Table 2 Focus Group Participants' Descriptions of Online Shopping

Goal-Oriented Shopping	Experiential Shopping
Accomplishment	Enjoyment
Going to specific site	Surfing/trying new sites
Looking for specific product	Looking for new things
Saving time	Killing time
Having a purpose in mind	Looking for ideas
Making repeat purchases	Checking favorite sites regularly
Finding the best price for a specific item	Bargain hunting for what's on sale

Note. Source: Wolfinbarger & Gilly (2001).

Table 3 Focus Group Participants' Desires When Shopping Online

Goal-Oriented Shopping	Experiential Shopping
To get in and out quickly (fewest clicks)	A welcoming site that draws me in
To do it myself	To interact with other consumers
To not waste time	To have lots of choices
To get immediate response to questions	To browse sites related to my hobby
Ease of use	A unique experience

Note. Source: Wolfinbarger & Gilly (2001).

Table 4 Focus Group Participants' Descriptors of Shopping Online

Freedom and Control	Fun
Control what information I receive	Read reviews (but don't believe them)
No salespeople	Get drawn in
No lines/crowds	Excitement of bidding
Only brands/sites I know	Window shopping
Can come back anytime/delay purchase	Am impulsive
Have options	Have to limit myself
Show me what I want	Surprises

Note. Source: Wolfinbarger & Gilly (2001).

informativeness, selection, and lack of sociality—with increased freedom and control. We discuss each of these attributes below.

Convenience and Accessibility

For consumers, online purchase is the ultimate in saving time and effort, which is made possible by the continuous, 24 hours a day/7 days a week/52 weeks a year accessibility of online stores. Many consumers shop at work and school, in part because of the availability of faster Internet connections. Overall buying effort is also reduced online because shoppers do not need to worry about their appearance; in fact, as one Earthlink billboard proclaimed, they can even "Shop Naked."

Although making an online transaction is often more convenient, consumers are nevertheless unable to touch or try on items. Virtual model technologies are not yet adequate substitutes for trying things on, and zooming in on pictures and rotating them does not represent the texture and feel of most items. Consumers who need a product immediately—for example, a last-minute anniversary gift—cannot obtain it. The difficulty of experiencing some qualities of products online together with the inability to take immediate possession of goods mutes experiential shopping impulses in the online environment.

Selection

Online shoppers frequently offer increased product selection as a goal-directed reason to shop online. Some online buyers have to drive long distances to get to products they need or want; for them, buying tools on the Internet rather than driving an hour to Sears is often preferable. Specialty products and sizes can be efficiently offered on the Internet, where geographically dispersed customers can be aggregated and thus profitably served. For example, Eddie Bauer offers petite and large sizes through their Web site, but not their stores, because of insufficient demand for these sizes in the stores' geographic markets.

Online shoppers look to Web sites to extend the inventory available locally. This extended inventory and selection can be offered in-store through kiosks; while shopping at REI, consumers can access the company's Web site and get an extended selection of specialized products for hiking, mountain climbing, rock climbing, and biking that are not in stock in the store (Schwartz, 2001).

Availability of Information

Availability of information is an important resource cited by consumers using the Internet for specific product searches; the availability and depth of information that allow consumers to feel comfortable that they are making good purchase decisions are reasons that many buyers view online search and purchase as a goal-directed activity. The cost of searching in terms of time spent is dramatically reduced online. Improvements over bricks-and-mortar stores include the reduction of irrelevant information, better information organization, and useful information processing aids.

Another aspect of information important to customers is the interactivity of some e-commerce sites. For example, making plane reservations online gives buyers the feeling that they more fully control their investigation of options than they do offline. Information gotten online can be printed out and saved, unlike that gotten in a phone conversation with a travel agent. In addition to cost savings for the airlines, the ability of consumers to easily access comparative information for different times of day and various airlines is a reason that penetration of airline ticket sales online is expected to grow; in fact, Boston Consulting Group reports that by 2005, fully one third of consumers anticipate spending more than half of their travel budgets on trips that were arranged online (Silverstein, Stanger, & Abdlemissih, 2001). Price comparison sites such as Dealtime.com and MySimon.com enable consumers to choose specific products and then get listings comparing prices at several different online e-tailers. Again, using these sites increases consumer perceptions of control over purchase transactions.

Lack of Sociality

It is interesting that online buyers are quite positive about the relative lack of social interaction experienced while buying online. Salespeople, spouses, crowds, and lines don't exist at e-tailing Web sites. Moreover, the ability to locate the information they need and to complete their transaction without having to interact with a human being increases their sense of freedom and control. Anonymity represents another positive aspect of feeling alone while shopping at Web sites; some online buyers visit sites of stores where they might be intimidated or embarrassed to shop offline, such as Victoria's Secret or Armani.

Consumers appreciate the absence of retail workers online because salespeople are often perceived to be unhelpful or uninformed. In addition, salespeople may pressure buyers or make them feel obligated, giving them a sense of less control over the buying situation. In fact, some consumers even prefer to avoid helpful salespeople because they dislike feeling obligated to purchase. As with other goal-directed themes, the lack of people to interact with online, including a spouse, who might hurry them along or wander off on unplanned buying forays, is associated with freedom. Despite the fact that online consumers largely like being able to avoid people online, they do sometimes require help and want to interact with a human being when they do. E-mail assistance helps satisfy the online need for help, but it is sometimes perceived as being too slow or as not really useful for answering questions. E-commerce providers can improve their sites by combining the properties of impersonal mass communication with more personal communication, utilizing e-mail, chat rooms, or telephones. Users, however, want options as to what medium they use to contact a company and expect responsive service whenever it is needed.

The Experiential Motives of Online Shoppers

Although most online shopping trips are goal focused, as many as one third are experiential (Wolfinbarger & Gilly, 2001). Experiential shopping occurs at Web sites when consumers are browsing largely to be entertained; experiential shopping is associated with increased impulse spending in both online and offline shopping. Experiential buyers browse for three main reasons: auctions, ongoing hobby-type activities, and bargain hunting (as compared to price comparison for a specific preplanned purchase). When asked about browsing online, experiential consumers mention auction activities and ongoing hobby-type searches most frequently. Auctions such as eBay have greater "stickiness" than other e-commerce sites; visits to eBay consistently average an astounding 45 minutes to more than an hour, according to weekly NielsenNetratings. Auctions present several experiential benefits. Products on auction sites change often and many of the products are unusual, unique, or collectible. Bidding for products introduces an element of risk and gaming. Good surprises, unique products, and excitement are all experiential benefits associated with auction sites. More than two thirds of auction participants shop with experiential motivations.

Experiential online shoppers often report hobbies that they actively pursue while online. The most natural hobby is computers and software, but online environments support a wide range of hobby-oriented communities and e-commerce sites as well, such as for campers, snowboarders, Barbie collectors, Texas Instrument calculator enthusiasts, and bibliophiles. These enthusiasts check sites related to their hobbies regularly, looking for new items and considering planned future purchases. Like auction participants, hobbyists are more likely than nonhobbyists to be interested in surprises, new features and items, and online community. A third activity that results in experiential browsing online is looking for deals. While goal-directed shoppers often seek comparative price information, they do so to comparison shop for a specific product rather than to hunt for previously unplanned purchases that happen to be bargains. A relationship between deal seeking and experiential motivations has been found in offline contexts as well.

Multichannel Shopping

It is important that most customers are now engaging in multichannel shopping. Those consumers who shop in more than one channel tend to spend more than those who shop only in the offline store. Many shoppers appreciate the advantages of multichannel shopping: going to an offline store where they can examine and try merchandise before buying online, or doing their information search and pricing online and then purchasing in stores. Industry analysts and consultants such as Jupiter Media Metrix are working to conceptualize valuation models that include the value of a Web site in creating offline as well as online purchases. Companies that focus exclusively on online purchases in calculating the return on investment in their Web sites may be seriously undervaluing their online efforts (Pastore, 2001b).

Although consumers like to shop online when they are goal-oriented, they shop offline when they want to physically examine items and enjoy the stimulation of being out. Consumers can be goal focused when they are time starved and experiential when they have the time to enjoy the experience of shopping. Consumer desires for multisensory experiences are much more likely to be met offline; when consumers desire these experiences, they are unlikely to shop online. For this reason, online shopping is unlikely to outweigh offline shopping for most product categories for the foreseeable future. Those product categories that are most likely to reach the highest online penetration rates are the following: (a) Products for which rich information is helpful in purchase, (b) commodities like books and CDs, (c) product categories in which extensive selection is important to a sizable niche of consumers, and (d) any product that can be downloaded, for example software and event tickets.

Meeting the Needs of Goal-Focused Online Shoppers

Given the fact that most online consumers are goal-oriented, marketers must design Web-based features that facilitate the efficiency these shoppers demand. Only experiential consumers desire community and nonproduct content, and many e-commerce sites may find that they have too few experiential consumers to merit special features designed for them. Even luxury brands like Tiffany's are finding that online consumers are not browsing; they are looking for an easy-to-use site that facilitates product search and easy transactions. A Forrester study found that two thirds of all e-commerce transactions that are begun are never completed, mostly because shoppers cannot find the information they need to complete the purchase (Haney, 2000). E-tailers must largely learn to ignore the siren call to implement attractive, technologically innovative, and clever features that nevertheless do not make a shopper's experience easier or more efficient. Thus, companies that implement experiential features in their site may be overlooking the needs of goal-oriented customers. Web site developers need to accept that online consumers will become loyal customers when they are well served (Reichheld & Schefter, 2000). For most online consumers, being well served means finding the product selection and information they desire, making the transaction easily, and receiving the product in a timely fashion. The importance goal-oriented consumers place on freedom from salespeople, spouses, and crowds at least partially explains why many efforts designed to integrate community with commercial sites have met with limited success.

Meeting the Needs of Experiential Consumers

Although the majority of online consumers tend to be goal oriented, experiential browsing behavior is desirable for online marketers because it often results in increased impulse purchases (Solomon, 1999). As young Internet users who have grown up on the net become independent consumers and broadband use becomes more widespread, experiential benefits may become desirable to more consumers. Before emphasizing such benefits, however, marketers must first make sure that the Web site is easy to search and navigate. Moreover, analysis must establish that the site will draw a sizable proportion of experiential shoppers to the product category. As they do at retail stores (Babin, Darden, & Griffen, 1994), experiential shoppers spend more time at sites and visit more often. Thus, using clickstream data, companies can estimate what proportion of visits are experiential by identifying consumers who visit more often, access more pages, and spend more time at the Web site (Hoffman & Novak, 1996).

SEGMENTING ONLINE CONSUMERS

Up until now, we have characterized shopping moods, rather than shoppers, and suggested that consumers who are goal oriented are more likely to shop online. However, it is useful to segment and describe the shoppers themselves as well. Forrester's *Technographics,* which is widely published, segments consumers into 10 segments based on attitude toward technology (positive or negative), income (high or low), and motivation to use technology (career, family, or entertainment). Although higher income ($40,000+) consumers are more likely to be technology optimists, a surprising 40% of the higher income category is made up of technology pessimists. The 10 groups can be categorized into three larger groups. Early adopters are characterized as Fast Forwards, New Age Nurturers, and Mouse Potatoes (29%) The mainstream group is as much as 2 years behind the early adopters; they are Techno Strivers, Digital Hopefuls, Gadget Grabbers, Handshakers, Traditionalists, and Media Junkies (43%). Laggards are sidelined citizens (28%).

Among the early adopters are the earliest of adopters, the Fast Forwards. The earliest and heaviest Internet shoppers, they are high income, motivated to use the Internet primarily by career needs, and likely to own many consumer technology products. They are largely efficient, focused shoppers and are not particularly price oriented. A second early adopter group is the New Age Nurturers. This group is also high income, but compared to the Fast Forwards, they're more likely to be motivated to use the Internet to facilitate their interests in family and community. New Age Nurturers are more price oriented than Fast Forwards. Forrester believes that New Age Nurturers are more likely to be credible opinion leaders for mainstream users and thus are more influential to later adopters than are the more innovative Fast Forwards. The growing popularity of clicks-and-mortar businesses is partially due to the influence of New Age Nurturers. A final segment is also included among the early adopters, the Mouse Potatoes. Mouse Potatoes use technology for fun, enjoy sports sites and adult entertainment, and play online games. They are more likely to be male, less likely than other early adopters to have children, and have a median age of 40. Like other early adopters, they are high income, but instead of being deliberate shoppers like the Fast Forwards, they tend to buy on impulse.

The second group of segments expected to adopt online shopping are called the Mainstream. As mentioned in the chapter's introduction, the next generation of online shoppers is anticipated to be more representative of middle-class America, with nearly half of new online shoppers' household incomes ranging from $30,000 to $75,000 between 2002 and 2006. The mainstream includes consumers who are high-income technology pessimists and low-income technology optimists. Among the high-income pessimists are the career-oriented Handshakers, the family and community-oriented Traditionalists, and the entertainment-oriented Media Junkies, who like fun, but nevertheless aren't very interested in technology. Mary Modahl writes:

> Demographically, the [technology] pessimists look just like the optimists—middle aged, well educated, family householders. High-income pessimists like to drive Jeeps, Volvos, and Toyotas, just like high-income optimists. What's more, high-income consumers share tastes in media. Both like newsmagazine shows and major sporting and entertainment events (2000, p. 49).

Thus, high-income pessimists are much like high-income optimists except for their attitudes toward technology.

The low-income technology optimists include the Techno Strivers, Gadget Grabbers, and Digital Hopefuls. The Techno Strivers are career oriented whereas the gadget grabbers are fun oriented. Techno Strivers and Gadget Grabbers are demographically diverse and include strong representation of both males and females, almost one quarter are minorities, and they engage in online activities. They don't yet earn much, mostly because they're just starting out. They tend to be single, college educated, and white collar. The last group of low-income optimists is Digital Hopefuls, a very large group of family-oriented retirees using the Internet to stay in touch. In fact, although retirees do not yet participate in the Internet in proportion to their number in the population, those who do participate spend more time online per week than do people from any other age group. The last segment of consumers is not likely to become Internet savvy at all. They are Sidelined Citizens. As Modahl summarizes:

> These consumers are pessimistic about technology. Even if they could afford a home PC, most probably wouldn't buy one. And if they had a PC, it is unlikely that these consumers would use it for online shopping . . . [F]ewer than 50 percent of Sidelined Citizens use the automated teller machines at their bank for making deposits and withdrawals. That technology is free, and it has been around for more than 20 years. (2000, p. 68)

RELATIONSHIP OF ATTRIBUTES OF ONLINE SHOPPING TO CONSUMER JUDGEMENTS OF QUALITY AND SATISFACTION

We have established that online shopping tends to be solitary and purposive rather than interpersonal and social by nature. But what do consumers really want from their online shopping experiences? What attributes are most important in their judgments of quality, satisfaction, and loyalty? These questions are important; buyers' perceptions of quality are likely to play a role in "e-loyalty" and profitability (Reichheld & Schefter, 2000).

Along with an increase in consumer experience online comes an increase in expectations of online businesses. Market research and consulting firms have been particularly interested in measuring e-commerce quality. A well-known example is Bizrate.com: measurements are based on intercept surveys customers receive immediately after making an online purchase. Bizrate measures ease of ordering, product selection, product information, price, on-time delivery, product representation, customer support, privacy policies, and shipping and handling. Scholars have studied a variety of factors, including usability, information content, selection of products, fulfillment, customer service, privacy/security, and Web site atmospherics. Following is a review of each of these factors and their impact on consumer evaluations of Web sites.

Usability

Before the Internet was widely available, information systems scholars and practitioners studied usability as it applied to computer systems, usually in work rather than home settings (Cooper, 2000). In general, usability studies examine user/technology interfaces in order to improve design so that it is more user oriented (Lohse & Spiller, 1998; Nielsen, 1993, 2000). Applied to Web sites, the term usability means intuitive navigation, a good search function, and easy checkout. Usability elements affect the ease of use and amount of mental effort required to browse and buy at an e-tailing site; a Web site that is difficult to use deters online consumers, especially less-experienced Internet users (Lohse & Spiller, 1999; Montoya-Weiss, Voss, & Grewal, 2000). Slow downloading raises the likelihood that customers will abandon the site (Nielsen, 2000; Dallaert & Kahn, 1999). Although consumer technology affects waiting time, consumers know which sites are quicker relative to the computer they normally use. E-tailing sites should test their Web site designs with 10–15 consumers directed to carry out specific instructions regarding the site. Interviewing consumers while they are carrying out these instructions will identify the major usability problems that can occur at e-tailing Web sites (Nielsen, 2000).

Information Content

An important advantage of online shopping is the availability of information, including the reduction in search effort and time for products and product-related information. The easy availability of in-depth information, as long as it is well organized and easy to access, is an important reason consumers shop online (Alba, et al., 1997; Ariel, 2000; Lynch & Ariely, 2000; Van den Poel & Leunis, 1999; Wolfinbarger & Gilly, 2000). As we have already discussed, such information does not just simply “replace” a salesperson; many online buyers report that they appreciate obtaining, and prefer to obtain, information directly, without having to go through a salesperson or retail worker. In addition, online information is perceived as superior to catalog shopping, as more information is available online and queries can be answered (Ariel, 2000; Van den Poel & Leunis, 1999; Venkatesh, 1998).

The ability of online consumers to search and compare price and quality information across Web sites increases their satisfaction with the shopping experience and the purchased product. As well, the belief that a Web site has useful information is associated with intentions to revisit the Web site and to repurchase. Further, interactive control of the content, order, and duration of product-relevant information improves consumers’ ability to integrate, remember, and use information (Lynch & Ariely, 2000).

Fulfillment/Reliability

Reliability, a key services quality concept typically rated as the most important to consumers (Zeithaml, Berry, & Parasuraman, 1993), is related to consumer perceptions of online transaction quality. Online e-tailers need to convince buyers that they can deliver products and services reliably and at the promised level of quality. In e-tailing, qualities associated with reliability are related to fulfillment and include on-time delivery, having products in stock, and effectively portraying the product on the Web site so that what is received is what the consumer expected to receive. Consumer ratings of fulfillment together with those of security/privacy strongly predict consumer trust of an individual Web site (Jarvenpaa & Tractinsky, 1999; Palmer, Bailey, & Faraj, 1999).

Customer Service

In the retailing literature, the level of service received by customers is frequently noted as a component of store image (Louviere & Johnson, 1990; Reardon, Miller, & Coe, 1995). In their measurement of services quality, Parasuraman, Zeithaml, and Berry (1988) included responsiveness, which is defined as the willingness to help customers, and empathy, defined as caring, individualized attention. In addition to the empathy and responsiveness of live customer service personnel, there are unique aspects of customer service in online retailing. Fast response to e-mail inquiries is an important attribute of online customer service. As well, ease of returning items is especially important to consumers in online environments and is thus a factor in their rating of online customer service. Customer service is particularly important when an online customer has a problem; because of the intangible nature of Web sites, customer service is a key in reassuring consumers that the company has real people who will solve their real problems (Wolfinbarger & Gilly, 2001).

Product Selection

Because researchers of computer-mediated environments until recently have focused on Internet use generally, rather than e-tailers specifically, scholars have not addressed online selection as extensively as other elements of the online shopping experience. As earlier reviewed in this chapter, increased product selection is an important motivation for consumers to buy online rather than offline. Moreover, because selection has been found to be consistently important in the literature concerning bricks-and-mortar retailers, it is expected that selection will be important to customers of online retailers as well (e.g., Reardon, Miller, & Coe, 1995; Samli, Kelly, & Hunt, 1998).

Privacy/Security

In the online arena, a sense of assurance is likely to be defined by consumers as adequate security and privacy (Culnan & Armstrong, 1999; Culnan, 1993; Hoffman, Novak, & Peralta, 1999). Security risk is the risk that credit card information is not safe. A second online risk is privacy risk, which involves the possibility that personal information shared with a Web site will be used for a secondary purpose without the consent of consumers (Culnan, 1993; Hoffman, Novak, & Peralta, 1999). Online shoppers largely understand that detailed information can be collected online concerning their search and purchase behaviors (Hoffman, Novak, & Peralta, 1999). In fact, in a Harris Interactive Study, fully 45% of people in the United States said they believed that online buying is a threat to privacy. Moreover, over two thirds of Internet users cite three concerns regarding online privacy: (a) Companies will provide their information to other companies without permission; (b) their transactions with companies may not be secure; and (c) hackers could steal their personal data from companies (Emarketer.com, 2002).

People who buy online are slightly more likely to trust Web sites than others (Ulsaner, 2000). Because consumers do not have the expertise to determine whether a Web site is safe, consumers see symbols such as the padlock-and-key icon or the TRUSTe seal as evidence for site security (Friedman, Kahn, & Howe, 2000); however, many consumers remain distrustful of sites despite their posted privacy policies and use of the key icon. The success of retailers already well trusted in the offline environment can be attributed partially to the belief that these companies are more likely than unknown companies to protect customer privacy and credit card security. Appearance and functionality of the Web site is related to consumers' perceptions of privacy and security risks. As consumers in offline environments have shown, they will rely on available tangible cues when they have uncertainty about a retailer (Doney & Canon, 1997).

Web Site Atmosphere

A more inviting Web site atmosphere may draw more experiential shoppers. More experiential shopping is desirable for e-tailers, as it has been associated with playfulness, positive feelings, and increased time spent online (Novak, Hoffman, & Yung, 2000). As well, in offline environments, experiential shopping has been associated with more impulsive and increased spending (Babin, Dardin, & Griffen, 1994). Hoffman and Novak (1996) write extensively on the possibility for online "flow," which involves loss of self-focus, intense involvement during an Internet session, and loss of a sense of time. It is important that the authors suggest that flow underlies a "compelling online customer experience." Marketers have speculated that creating "a compelling online experience" is likely to be strategically important, given that increasingly transparent information environments have been created on the Internet (Häubl & Trifts, 2000; Lynch & Ariely, 2000). However, it is unclear whether atmosphere is part of that compelling experience or even whether it is possible to create an atmosphere in an online store; it may be that Web site usability is more important in creating a compelling experience than is atmosphere. Moreover, the predominance of goal focused shopping online casts doubt on the importance of Web site atmospherics; currently, the effect of an attractive and pleasing Web site design along with various personalization techniques is not yet fully understood.

In offline retailing, consumer evaluation of a store's appearance and atmosphere is one of several predictors of store image ratings (cf. Doney & Canon, 1997). The graphic style of a Web site is similar to the interior and exterior appearance of a bricks-and-mortar store; as with offline stores, the style and appearance of a Web site provide messages about the positioning of the site. Nevertheless, it is unclear how atmospheric elements can be created that are experiential but which do not interfere with the needs of goal-oriented consumers. To date, consumers have consistently expressed dissatisfaction with experiential graphic features that are attractive, but which slow downloads and impede usability. The splash sequences widely used by Web sites in 1998 and 1999 that impeded entry into a Web site are an excellent example of an attractive design feature that consumers largely rejected (Nielsen, 2000).

One element that may contribute to an inviting Web site atmosphere is appropriate personalization. Personalization is providing "users with what they want or need without requiring them to ask for it explicitly" (Mulvenna, Anand, & Buchner, 2000, p. 123). In the online setting, personalization includes recommendation systems based on collaborative filtering or observational techniques, customization, and adaptive Web sites (Sipiliopoulou, 2000). However, a majority of Web users find online solicitations to be intrusive, which suggests that e-tailers must walk a fine line between personalization and personal intrusion. In addition to being related to the atmosphere of a Web site, personalization is also related to usability, as personalization can result in making the site more intuitive and easy to use. Personalization is seen as positive to users when it includes features that increase the user's sense of control and freedom, such as order tracking, purchase histories, saving information for quicker transactions during future sessions, and opt-in e-mail notification of new products and special deals. Personalization is seen as negative, however, when the result is unsolicited offers or if users feel less anonymous. These intrusive features decrease users' perceived control and freedom (Wolfinbarger & Gilly, 2001).

Measuring Online Quality From the Consumers Point of View

Scholars and researchers have recently begun to model and measure the dimensions of quality of the online buying experience. These research efforts vary in focus and, thus, measure differing aspects of e-commerce quality. For example, WEBQUAL (Loiacono, Watson, & Goodhue, 2002) measures various dimensions of the Web site interface or design, including information fit-to-task, interactivity, trust, response time, ease of use, entertainment, consistent image, online completeness, and substitutability for offline channels. Symananski and Hise (2000) study “e-satisfaction” and conclude that online convenience, merchandising (product offerings and information), site design, and financial security all contribute to customer satisfaction. Left out of both WEBQUAL and Symanski’s and Hise’s e-satisfaction, however, are customer experiences with fulfillment and customer service.

Two recent scale efforts have some commonalities both in their approach and in their outcomes, eTailQ (Wolfinbarger & Gilly, 2002, in press) and eSQ (Zeithaml, Parasuraman, & Malhotra, 2000, 2002). Both efforts include measurements of the entire process of making an online purchase, from looking for information at the Web site, to making a transaction, to product delivery and customer service. Following is a more in-depth description of eSQ and eTailQ.

eTailQ: Dimensionalizing, Measuring, and Predicting Online Quality

eTailQ is based on a three-step development effort, including (a) focus groups to uncover attributes of importance, (b) an intermediate step in which online consumers categorize attributes into meaningful groups, and (c) an online survey utilizing consumers from Harris Interactive’s online research panel. The developers of eTailQ conclude that across auction and e-tailing settings, browsing versus buying purchase situations, and a variety of product categories, four higher level constructs predict customer judgments of online quality (see Table 5): Web site design (conceived of broadly as information availability, usability, product selection, and appropriate personalization), fulfillment/reliability (receiving the product ordered within the time promised), security/privacy (security of credit card payments and privacy of shared information), and customer service (responding to e-mails, handling returns, answering questions, and being responsive and helpful).

Table 5 eTailQ Items (short version)

Web site design
The Web site provides in-depth information.
The site doesn’t waste my time.
It is quick and easy to complete a transaction at this Web site.
The level of personalization at this site is about right, not too much or too little.
This Web site has good selection.
Reliability/fulfillment
The product that came was represented accurately by the Web site.
You get what you order from this Web site.
The product is delivered by the time promised by the company.
Privacy/security
I feel my privacy is protected at this site.
I feel safe in my transactions with this Web site.
This Web site has adequate security features.
Customer service
The company is willing and ready to respond to customer needs.
When you have a problem, the Web site shows a sincere interest in solving it.
Inquiries are answered promptly.

Graphic design was not as important to online consumers as other Web site design features and was not included in the final eTailQ measurement instrument. The factors having the strongest impact on both customer satisfaction and quality judgments are Web site design and fulfillment. Customer ratings of overall quality predict that about half of the variance in customer satisfaction is related to their most recent online purchase, while pricing explains an additional 25%.

Perhaps most surprising is the role of security/privacy, which does not predict customers’ rating of Web site quality, except among the most frequent buyers at the site. It is important that security/privacy is strongly correlated with consumers’ evaluations of Web site design; thus, it appears that inferences of security/privacy are largely made from Web site design when consumers do not yet have experiences with a Web site. In focus groups, online consumers say that they have difficulty judging the privacy and security of a site, even after checking that the site is secure when making transactions and after reading a statement of privacy that in their minds was legalistic. Thus, customers assess the professional look and feel of a Web site, the functionality of a Web site, and the company reputation, and they extrapolate from that to judge privacy/security. Usable and professional sites require greater resources and investment and thus inspire greater consumer trust. Similarly, in a brick-and-mortar context, consumer trust is affected by a seller’s perceived investment in a facility (Doney & Canon, 1997).

Although Web site design and fulfillment are about equally predictive of customer ratings of Web site quality, loyalty and repurchase intentions are most strongly related to Web site design elements. This is likely to be because of the fact that even when the product and the delivery are satisfactory, if the Web site experience is not, customers are much less apt to return to the Web site for future purchases. The experience consumers have at a Web site takes on added strategic importance when you consider that many more consumers purchase products offline as a result of online information search than actually buy online; these online shoppers who are not online buyers are likely to have experiences with Web site design but not with the other three quality factors—fulfillment/reliability, security/privacy, and Web site customer service.

eSQ: Conceptualizing Online Experiences as Services

The eSQ scale was developed through a three-stage process using exploratory focus groups and two phases of survey data collection and analysis (Zeithaml, Parasuraman, & Malhotra, 2002a). eSQ suggests that four dimensions form “core” service—efficiency (ease of Web site use), fulfillment, privacy, and technical reliability. Efficiency, fulfillment, and privacy are very similar to eTailQ’s Web site design, fulfillment/reliability, and privacy/security dimensions, respectively. The similarity in findings concerning these three dimensions across two research efforts provides support for their importance in predicting quality. The fourth core dimension, technical reliability, is not included in eTailQ and refers to correct technical functioning and availability of the Web site.

Rather than view customer service as a core element of a typical online purchase experience, as does eTailQ, eSQ suggests that customer service comes into play only when a customer problem occurs. The customer service scale includes three elements of "recovery" service—responsiveness, compensation, and contact. Responsiveness is defined as "the ability of e-tailers to provide appropriate information to customers when a problem occurs, have mechanisms for handling returns, and providing online guarantees" (Zeithaml, Parasuraman, & Malhotra, 2002a, p. 15). Compensation involves crediting money to consumers and handling returns. Contact is the ability to speak live with customer service personnel, either through online chat rooms or on the phone.

The authors of eSQ conceptualize four potential gaps in online service quality (Parasurman, Ziethaml, & Malhotra, 2000): The *information gap* is the difference between what managers believe consumers desire in Web site design and operations and what customers actually want. If consumer needs and wants with regards to desired Web site attributes are not continuously monitored, an information gap will exist. A *design gap* occurs when a company with accurate knowledge fails to incorporate its findings into Web site design and supporting operations such as customer service and fulfillment. A *communications gap* occurs when more is promised at the Web site than can be delivered, often a result of competitive pressures. For example, quicker delivery dates than are actually possible may be promised by the marketing department when, in fact, these promises cannot be supported by the shipping department. A *fulfillment gap* is the sum total of all discrepancies between what customers wanted and expected and what they actually received. All three of the other gaps contribute to the fulfillment gap.

The development and conceptualization of online e-commerce quality is just now emerging, and all the proposed scales will benefit from further tests of their psychometric properties. The validity and reliability of the Bizrate.com online purchase satisfaction scale has not been established. Moreover, it is through repeated usage and improvement that scales such as WEBQUAL, eTailQ and eSQ will become most useful to e-commerce researchers and managers (Zeithaml, Parasuraman, & Malhotra, 2002b).

CONCLUSION: FUTURE OF ONLINE SHOPPING

Online shopping will continue to grow among mainstream segments in the United States in the next few years. At the same time, online shopping is growing in more advanced European markets, South Korea, Malaysia, Australia, Canada, and New Zealand. It is important to note that technology will impact the growth of online shopping. Broadband makes online shopping easier and more convenient; as well, broadband will enable the creation of more sophisticated Web sites with enhanced functionality, which will increase the comfort level of online buyers. Perhaps more experiential Web site design features will be preferred by customers as broadband diffuses more widely. Mobile commerce (m-commerce) may also open up new possibilities, allowing, for example, your local favorite restaurant to send special offers to your cell phone when you're in the neighborhood, or your dry cleaner to offer a new service—cell phone notification when your dry cleaning is ready. Marketers are also experimenting with interactive television (ITV), which could allow users to order furniture, jewelry, clothing, or other products that appear in regular TV programming. ITV could also be used to allow consumers to request additional information about advertised products. These e-commerce applications and others will be successful to the degree that they are compatible with consumer lifestyles, are responsive to consumer motivations for shopping using a particular technology, offer advantages over current methods consumers use to buy products, and are easy to use, requiring limited learning on the part of consumers.

GLOSSARY

Bricks and Mortar A physical space where consumers may shop for and purchase products; the terms is also sometimes used to describe retailers who own physical retail stores.

Clicks and Mortar A retailer that sells products both online and offline.

Community A virtual online gathering place for Internet users.

Content Any kind of information on a Web site; in an e-commerce context, however, "content" may refer both to information about products and to articles and outside information at Web sites that are indirectly product related; an example is the content at imdb.com (the Internet Movie Database), which is utilized by consumers to facilitate information search about movies and by Amazon.com to support sales of videotapes, DVDs, and CDs through links back to Amazon.

Experiential shopping Shopping primarily for fun rather than to make a preplanned purchase.

Fulfillment/reliability As seen from the consumer's point of view, fulfillment involves displaying and describing a product accurately on the Web so that what customers receive is what they thought they ordered, having the product in stock, and delivering the right product within the time frame promised.

Goal-directed shopping Shopping with a predetermined goal, usually to purchase or search for information for specific products; although such shopping may be enjoyable, the primary purpose of goal-directed shopping is product acquisition.

Haptics The science concerned with tactile senses; haptic devices allow users to experience the sensation of touch remotely.

Internal locus of control A personality characteristic marked by the tendency to be driven by internal goals and needs rather than external cues.

Multichannel shopping Shopping through two or more channels (online shopping, catalog shopping, and/or land-based shopping); for example, consumers may search for product information online and then purchase offline, or they may purchase in both the bricks-and-mortar and online stores of a particular retailer.

Personalization Any Web site strategy that involves customizing Web site appearance and functionality and/or making offers to consumers based on past behavior at a Web site.

Pureplay A company that only has an online presence without a bricks-and-mortar counterpart.

Sociality The state or quality of being sociable; sociability.

Technology optimism The belief that technology improves living; an orientation that predicts acceptance of new technologies.

Web site design All elements of the consumer's experience at the Web site (except for customer service), including navigation, information search, order processing, appropriate personalization, and product selection.

Wired lifestyle A lifestyle wherein individuals use the Internet for many purposes and thus naturally turn to the Internet to search for product information and to buy products and services.

CROSS REFERENCES

See *Click-and-Brick Electronic Commerce; Data Mining in E-Commerce; Electronic Commerce and Electronic Business; Intelligent Agents; Internet Literacy; Marketing Communication Strategies; Online Communities; Personalization and Customization Technologies; Web Site Design.*

REFERENCES

Alba, J., Lynch, J., Weitz, B., Janiszewksi, C., Lutz, R., Sawyer, A., & Wood, S. (1997). Interactive home shopping: Consumer, retailer and manufacturer incentives to participate in electronic marketplaces. *Journal of Marketing, 61*(3), 38–53.

Ariel, D. (2000). Controlling the information flow: Effects on consumers' decision making and preferences. *Journal of Consumer Research, 27*(2), 233–248.

Babin, B. J., Darden, W. R., & Griffen, M. (1994). Work and/or fun: Measuring hedonic and utilitarian shopping value. *Journal of Consumer Research, 20,* 644–656.

Bellman, S., Lohse, G., & Johnson, E. (1999). Predictors of online buying behavior. *Communications of the ACM, 42*(12), 32–38.

Berry, L., Parasuraman, A., & Zeithaml, V. (1993). More on improving service quality measurement. *Journal of Retailing,* Spring, 141–147.

Center for Media Research (2001). *Worldwide online shoppers*. Retrieved September 28, 2002, from http://www.mediapost.com/research/cfmr_briefArchive.cfm?s = 95879

Cooper, A. (2000). *The inmates are running the asylum: Why high-tech products drive us crazy and how to restore the sanity.* Indianapolis, IN: SAMS.

Culnan, M. J. (1993). How did they get my name? An exploratory investigation of consumer attitudes toward secondary information use. *MIS Quarterly, 17*(3), 341–363.

Culnan, M. J., & Armstrong, P. K. (1999). Information privacy concerns, procedural fairness and impersonal trust: An empirical investigation. *Organization Science, 10*(1), 104–115.

Dallaert, B., & Kahn, B. E. (1999). How tolerable is delay? Consumers' evaluations of internet Web sites after waiting. *Journal of Interactive Marketing, 13*(1), 41–54.

Doney, P. M., & Canon, J. P. (1997). An examination of the nature of trust in buyer-seller relationships. *Journal of Marketing, 61,* 35–51.

Emarketer.com (2002). *Consumers' concern with company security.* Retrieved February 22, 2002, from http://www.emarketer.com/estatnews/estats/ecommerce_b2c/2002022_harris.html

Forrester Research, Inc. (2002). *December shopping up from last year in spite of rough economy, according to the Forrester Research Online Retail Index*. Retrieved January 24, 2002, from http://www.forrester.com/ER/Press/Release/0,1769,678,00.html

Friedman, B., Kahn, P. H., Jr., & Howe, D. (2000), Trust online. *Communications of the ACM, 43,* 34–40.

Haney, K. (2000). *Utilize technology to keep shoppers.* Retrieved June 27, 2002, from http://www.digitrends.net/digitrends/dtonline/qa/qu062700.shtml

Häubl, G., & Trifts, V. (2000). Consumer decision making in online shopping environments: The effects of decision aids. *Marketing Science, 19*(1), 4–21.

Hoffman, D. L., & Novak, T. P. (1996). Marketing in hypermedia computer-mediated environments: Conceptual foundations. *Journal of Marketing, 60,* 50–68.

Hoffman, D. L., Novak, T. P., & Peralta, M. A. (1999). Building consumer trust online. *Communications of the ACM, 42*(4), 80–85.

Hoffman, D. L., Novak, T. P., & A. E. Schlosser (2000). *Consumer control in online environments. Working Paper, Vanderbilt University.* Retrieved June 6, 2002 from http://http://ecommerce.vanderbilt.edu/research/manuscripts/index.htm

Hoque, A. Y., & Lohse, G. L. (1999). An information search cost perspective for designing interfaces for electronic commerce. *Journal of Marketing Research, 36,* 387–394.

Jarvenpaa, S. L., & Tractinsky, N. (1999). Consumer trust in an Internet store: A cross-cultural validation. *Journal of Computer Mediated Communication,* 5 (2). Retrieved February 22, 2000 from http://www.ascusc.org/jcmc/vol5/issue2/jarvenpaa.html

Klein, L. (1998). Evaluating the potential of interactive media through a new lens: Search versus experience goods. *Journal of Business Research, 41,* 195–203.

Li, H., Kuo, C., & Russell, M. G. (1999). The impact of perceived channel utilities, shopping orientation and demographics on consumer's online behavior, *Journal of Computer-Mediated Communication,* 5(2), December. Retrieved February 9, 2000 from www.jcmc/vol5/issue2/hairong.html

Lohse, G., Bellman, S., & Johnson, E. (2000). Consumer buying behavior on the Internet: Findings from panel data. *Journal of Interactive Marketing, 14,* 15–29.

Lohse, G. L., & Spiller, P. (1998). Electronic shopping. *Communications of the ACM, 41,* 81–88.

Lohse, G. L., & Spiller, P. (1999). Internet retail store design: How the user interface influences traffic and sales. *Journal of Interactive Marketing, 14,* 15–29.

Loiacono, E., Watson, R. T., & Goodhue, D. L. (2002). WEBQUAL: A measure of Website quality. In K. R. Evans & L. K. Scheer (Eds), *2002 Winter educators' conference: Marketing theory and applications,* Vol. 13, 432–438, Chicago: American Marketing Association.

Louviere, J. J., & Johnson, R. D. (1990). Reliability and validity of the brand-anchored conjoint approach to measuring retailer images. *Journal of Retailing, 66,* 359–382.

Lynch, J. G., & Ariely, D. (2000). Wine online: search costs affect competition on price, quality and distribution. *Marketing Science, 19* (1), 83–103.

Modahl, M. (2000). *Now or never: How companies must change today to win the battle for Internet consumers*. New York: Harper Business.

Mulvenna, M. D., Anand, S. S., & Buchner, A. G (2000). Personalization on the net using Web mining. *Communications of the ACM, 4,* 123–125.

Nielsen, J. (1993). *Usability engineering*. Cambridge, MA: Academic Press.

Nielsen, J. (2000). *Designing Web usability: The practice of simplicity*. Indianapolis, IN: New Riders Publisher.

Novak, T. P., Hoffman, D. L., & Yung, Y. F. (2000). Measuring the customer experience in online environments: A structural modeling approach. *Marketing Science, 19*(1), 22–42.

Palmer, J. W., Bailey, J. P., & Faraj, S. (1999). The role of intermediaries in the development of trust on the WWW: The use and prominence of trusted third parties and privacy statements. *Journal of Computer Mediated Communication*. Retrieved February 9, 2000 from http://www.ascusc.org/jcmc/vol5/issue3/palmer.html

Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1988). SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. *Journal of Retailing, 6* (4), 12–40.

Pastore, M. (1999). *Affluent women buy on the Web*. Retrieved May 18, 2002, from http://cyberatlas.internet.com/markets/retailing/article/0,,6061_154071,00.html

Pastore, M. (2001a). *Where in the world is the best e-commerce*. Retrieved May 15, 2002, from http://cyberatlas.internet.com/big_picture/geographics/print/0,,5911_766411,00.html

Pastore, M. (2001b). *There's more to e-commerce than pnline profits*. Retrieved August 20, 2002, from http://asia.internet.com/asia-news/article/0,3916,161_868861,00.html

Pastore, M. (2002). *Despite economy, European e-commerce forecasts stand*. Retrieved January 10, 2002, from http://cyberatlas.internet.com/big_picture/geographics/article/0,,5911_952691,00.html

Reardon, H., Miller, C. E., & Coe, B. (1995). Applied scale development: Measurement of store image. *Journal of Applied Business Research, 11,* 85–93.

Reichheld, R., & Schefter, P. (2000). E-Loyalty: Your secret weapon on the Web. *Harvard Business Review, July-August,* 105–113.

Samli, A. C., Kelly, J. P., & Hunt, H. K. (1998). Improving the retail performance by contrasting management—and customer—perceived store images: A diagnostic tool for corrective action. *Journal of Business Research, 43,* 27–38.

Schwartz, E. I. (2001). *Digital Darwinism: 7 breakthrough business strategies for surviving in the cutthroat Web economy*. New York: Broadway Books.

Silverstein, M., Stanger P., & Abdelmessih, N. (2001). *The next chapter in business-to-consumer e-commerce: Advantage e-commerce*. Boston, MA: Boston Consulting Group.

Sipiliopoulou, M. (2000). Web usage mining or Web site evaluation. *Communication of the ACM, 42,* 127–134.

Solomon, K. (1999). *Revenge of the Bots. The industry standard.* Retrieved November 15, 2002, from http://www.thestandard.com/article/display/0,1151,7624,00.html

Symanski. D., & Hise, R. (2000). e-Satisfaction: An initial examination. *Journal of Retailing, 76*(3), 309–322.

Tedeschi, B. (2002). E-commerce report: Though there are fewer Internet users, experienced ones, particularly the middle aged, are increasingly shopping online. *The New York Times,* March 4, p. C7.

Ulsaner, E. M. (2000). Social capital and the net. *Communications of the ACM, 43,* 60–64.

Van den Poel, D., & Leunis, J. (1999). Consumer acceptance of the Internet as a channel of distribution. *Journal of Business Research, 45,* 249–256.

Venkatesh, A. (1998). Cybermarkets and consumer freedoms and identities. *European Journal of Marketing, 32,* 664–676.

Wolfinbarger, M. F., & Gilly, M. C. (2001). Shopping Online for freedom, control and fun. *California Management Review, 43,* 34–55.

Wolfinbarger, M. F., & Gilly, M. C., (2002). *.comQ: Conceptualizing, measuring and predicting e-tail quality. Working paper, report no. 02–100*. Boston, MA: Marketing Science Institute.

Wolfinbarger, M. F., & Gilly, M. C. (in press), eTailQ: Dimensionalizing, measuring and prediciting e-tail quality, *Journal of Retailing.*

Ziethaml, V., Berry, L., & Parasuraman, A. (1993). The nature and determinants of customer expectations of service. *Journal of Academy of Marketing Science, 21,* 1–12.

Zeithaml, V. A., Parasuraman, A., & Malhotra, A. (2000). *E-service quality: Definition, dimensions and conceptual model. Working paper series*. Cambridge, MA: Marketing Science Institute.

Zeithaml, V. A., Parasuraman, A., & Malhotra, A. (2002a). *An empirical examination of the service quality-value-loyalty chain in an electronic channel. Working paper*. Durham: University of North Carolina.

Zeithaml, V. A., Parasuraman, A., & Malhotra, A. (2002b). Service quality delivery through Web sites: A critical review of extant knowledge. *Journal of Academy of Marketing Science. 30* (Fall), 362–410.

Consumer-Oriented Electronic Commerce

Henry Chan, *The Hong Kong Polytechnic University, China*

INTRODUCTION

The beginning of the 21st century has marked the beginning of a global entity with a duplex (physical plus digital) market. Enabled by various Internet technologies, e-commerce forms a global marketspace, introducing numerous new services to consumers and creating a paradigm shift in many businesses. As complements to each other, the physical marketplace and the electronic or digital marketspace are leading us into a new generation of the economy. (See for example Rayport & Sviokla, 1995, and Rayport & Sviokla, 1994, about marketspace and marketplace.)

E-commerce is commonly related to the sale and purchase of goods and services by electronic or digital mechanisms. (See Kalakota & Whinston, 1997, for definitions of e-commerce from different perspectives.) It is usually associated with sales transactions on the Internet, although it is not confined to those activities. The predecessors of e-commerce are generally agreed to be electronic funds transfer and electronic data interchange services (Kalakota & Whinston, 1997; Zwass, n.d.), which are mainly concerned with interbusiness transactions. In contrast, today's e-commerce covers both business-oriented and consumer-oriented aspects. More important, e-commerce can now be supported over a public network (i.e., the Internet), thus facilitating its development and expansion.

As shown in Figure 1, there are two major types of e-commerce: consumer-oriented e-commerce and business-oriented e-commerce. Consumer-oriented e-commerce can be further divided into three types: business-to-consumer (B2C), consumer-to-consumer (C2C), and consumer-to-business (C2B) (Korper and Ellis, 2001; Turban, King, Lee, Warkentin, & Chung, 2002). Resembling a traditional retail business, B2C e-commerce is about the sale of goods or services to consumers through a business called an electronic retailer (e-retailer or e-tailer) on the Internet. "Brick-and-mortar" and "click-and-mortar" have become two popular terms to distinguish a traditional (physical) retailer from a hybrid (physical and electronic) retailer, respectively. Common features provided by an e-retailer include a search engine to facilitate searching for a product, an electronic catalog (e-catalog) to show product information, a virtual shopping cart to store selected goods, communication facilities to provide customer support and order handling functions to handle payment and product delivery (Schneider & Perry, 2001; Chan, Lee, Dillon, & Chang, 2001). For C2C e-commerce, both the buyers and sellers are consumers. Traditionally, used goods are often sold among consumers by means of classified advertisements. With the advent of Internet-based e-commerce, the auction model is becoming more popular. Emerging as a new mechanism, C2B e-commerce provides a buyer-centric and highly customized solution to fulfill a consumer's need. In contrast to consumer-oriented e-commerce, business-oriented e-commerce is concerned with commercial transactions between business partners. It can often be viewed as an essential back-end process to support consumer-oriented e-commerce. Business-oriented e-commerce generally covers supply chain management and associated transactions (Kalakota & Whinston, 1997).

In this chapter, an overview of e-commerce is presented with a focus on its consumer-oriented aspects. The organization of the rest of this chapter is as follows. The next section provides an overview of the consumer buying process. I then present different consumer-oriented e-commerce models and discuss the enabling technologies for consumer-oriented e-commerce. Next, I present two major developments in consumer-oriented e-commerce. The Conclusion provides a summary.

OVERVIEW OF THE CONSUMER BUYING PROCESS

Before discussing consumer-oriented e-commerce in detail, we return to the basics of consumer buying process. In fact, many consumer-oriented e-commerce models are designed to address this fundamental issue. In general, the consumer buying process consists of the following stages (Etzel, Walker, & Stanton, 2001):

Specify need(s),
Search for the alternatives,
Compare or evaluate the alternatives,

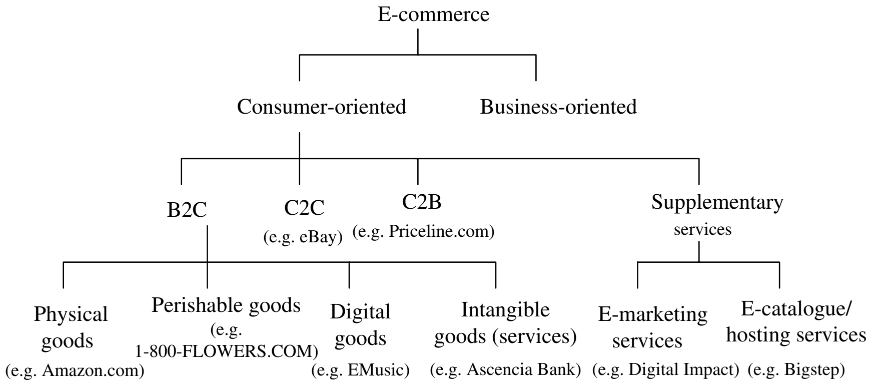


Figure 1: Different types of e-commerce.

Buy selected goods or services, and

Carry out other postpurchase activities.

It is best to explain this process with an example. (See also Etzel et al., 2001; Schneider & Perry, 2001; and Turban et al., 2002 for discussion on this process.) Suppose a consumer wants to learn more about e-commerce. Obviously, there are many alternatives. For example, he/she may buy a book (e.g., the *Internet Encyclopedia*), take a training course, or study for a formal degree. Hence, the first task is to define his/her needs more clearly. Let us assume that this consumer chooses to purchase a book on e-commerce. Next, he/she needs to search for the alternatives in the market by visiting bookstores. Having determined the alternatives, he/she needs to evaluate them. In the case of a book, the decision factors are likely to be price, experience of the authors and publishers, and whether the book is easy to read and understand. The consumer also needs to consider where to buy the book because the same book may be priced differently in various bookstores. After considering all these decision factors, he/she makes the decision and purchases the book from the selected bookstore. For nonstandard or semi-standard goods such as computers, further negotiations may be carried out to finalize the deal (e.g., to obtain a better service guarantee). Furthermore, for goods such as computers, there are also postpurchase activities such as technical support services.

In general, e-commerce can greatly facilitate the consumer buying process. In the physical commerce system, searching for information is often limited by various geographical constraints. For example, one can only visit the shops within a certain area. In other words, a consumer may not get sufficient information for comparison or evaluation purposes. In contrast, searching can be done more effectively in cyberspace. A consumer can also obtain more information to enable better decision making. E-commerce also facilitates evaluation, in particular, price comparison. To find the best available price of a product in the global marketspace, some price-comparison services are available on the Internet. For instance, a consumer may visit Best Web Buys (http://www.bestwebbuys.com) or mySimon (http://www.mysimon.com) to inquire about prices easily. Last, but not least, e-commerce also facilitates the purchase stage because consumers have better control (e.g., ease in changing the delivery information) and the process is automated. In the case of digital goods, consumers can obtain the goods instantly via the Internet. In summary, to buy something on the Internet, a consumer first searches for and evaluates the products through the use of directory services (e.g., Yahoo!; http://www.yahoo.com), search engines (e.g., Google; http://www.google.com) and price-comparison services (e.g., Best Web Buys or mySimon). Then he/she visits the virtual shop to complete the purchase. Finally, the purchased goods are delivered by physical or digital means. The buying process is also supported by other related services (e.g., payment services).

DIFFERENT CONSUMER-ORIENTED E-COMMERCE MODELS

The focus of this chapter is on consumer-oriented e-commerce in general and B2C e-commerce in particular. Inspired by some case studies presented in Chan, Lee, Dillon, and Chang (2001); Deitel, Deitel, and Steinbuhler (2001); Derfler et al. (2001); Korper and Ellis (2001); Schneider and Perry (2001); Turban et al. (2002); and the information obtained from the respective Web sites, various consumer-oriented e-commerce models and two supplementary services are discussed in the following sections and subsections.

B2C E-commerce

As noted earlier, B2C e-commerce involves the selling of goods or services by a business to consumers. It may be used to sell different types of goods including standard goods (e.g., books), perishable goods (e.g., flowers), digital goods (e.g., music), and intangible goods (e.g., banking services) as discussed in the following sections.

Selling Standard Goods: Amazon.com

Amazon.com (http://www.amazon.com) is an e-retailer established by Jeff Bezos for selling books. Because books are standard goods, they are suitable for sale through the Internet (Schneider & Perry, 2001); it makes little difference whether a book is bought from a physical or a cyber bookstore. In other words, the fact that consumers cannot check a book physically has little effect on their decision

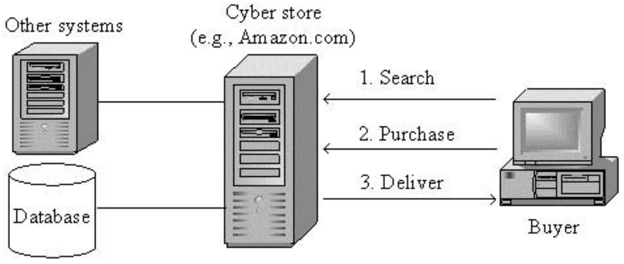


Figure 2: Cyberstore model.

to purchase it. From a business perspective, it is less costly to run a cyber bookstore than a physical bookstore.

Let us study how Amazon.com operates in the context of the consumer buying process (see Figure 2). When consumers visit Amazon.com, they can search for books through various mechanisms. In a physical bookstore, books are placed in labeled bookshelves according to their subjects. Using a similar arrangement, books are sorted by subject at Amazon.com to facilitate searching. Effectively, this can be viewed as a "translational" feature (i.e., to translate a physical feature into an equivalent electronic feature). In addition, Amazon.com also provides two "transformational" search features (i.e., they are not available in the traditional physical bookstore). The first feature is the search box, where a consumer can search for books by providing one or more keywords. Based on the keywords, a list of books will be displayed accordingly. The second feature is a proactive search feature called the recommendation system. By using data mining and other advanced computing techniques, Amazon.com can recommend books for consumers based on their interests (e.g., the books he/she has chosen before) and the buying behavior of other consumers. For example, if a buyer selects a book on e-commerce, Amazon.com will likely recommend other popular books on e-commerce to that person. In many cases, consumers will find the recommendations attractive and useful. Again, this transformational search feature greatly facilitates the search process.

Having found some books, the consumer can evaluate and compare them. Amazon.com offers some readers' ratings or scores and reviewer comments for reference. By reading the comments of other readers, the consumer can find out more about a book that he/she is interested in. Once the consumer has selected a book, he/she can store it in an electronic shopping cart and then choose the next book if required. Finally, the consumer can purchase the books in the shopping cart by going to the payment section. If registered with Amazon.com, the consumer may order through the 1-Click service by clicking the computer mouse once. Basically, the 1-Click service completes the order information (e.g., credit card number, delivery information, etc.), based on registered information. The shopping cart facility, together with this 1-Click service, greatly facilitates the purchase stage of the consumer buying process. Besides selling books, Amazon.com has now expanded "horizontally" to become a super e-store, selling many other goods (e.g., computers, kitchen appliances, software).

Like many other B2C Web sites, Amazon.com supports personalization and customization. Consumers registered with Amazon.com will see a personalized "hello" message with their names. Furthermore, Amazon.com allows its registered customers to build personalized stores. To reach as many consumers as possible, Amazon.com employs an affiliate scheme. This is a referral mechanism, in which an affiliate can link or refer its visitors to Amazon.com via Amazon.com's icon or even Amazon.com's search box. If a referred customer purchases something, the referring affiliate gets a commission. By using this incentive-based promotion service, Amazon.com can extend its coverage in cyberspace. In fact, the affiliate scheme has become a common and effective marketing technique adopted by many e-retailers.

Selling Perishable Goods: 1-800-FLOWERS.COM

1-800-FLOWERS.COM (http://www.1800flowers.com) sells flowers to consumers on the Internet. Similar to Amazon.com, it has the standard B2C features (e.g., a search engine, a shopping basket, communication and help facilities, payment functions) for facilitating a consumer buying process. It is also different from Amazon.com, however. Unlike books, flowers are perishable, so they need to be delivered in a timely manner and with great care. To cater to this important requirement, 1-800-FLOWERS.COM has many associated flower shops. The ordered flowers are delivered through these physical shops and other distribution outlets to the consumers. This example shows how e-commerce can work in conjunction with traditional commerce to satisfy a consumer need more effectively. Unlike books, flowers are seasonal products, and many people like to purchase flowers, for different occasions, on a regular basis. For example, a husband might like to buy flowers for his wife on her birthday each year. Furthermore, flowers are not standard goods because they can be packaged in a customized manner. Because of these special characteristics, 1-800-FLOWERS.COM provides many distinctive features compared with other e-retailers selling standard goods. First, the flowers are sorted by different means such as by occasion and price. Of course, 1-800-FLOWERS.COM also features some special and seasonal offers or promotional flowers in its e-catalog. Registered consumers (members) can enjoy a useful gift reminder service for the dates that they provide, and 1-800-FLOWERS.COM will remind them of those important days via e-mail. This is an effective way to generate steady revenue. In addition, members can also enjoy many other services, such as securely stored credit card information for subsequent orders, promotional e-mails, and Web-based order tracking.

Like Amazon.com, 1-800-FLOWERS.COM employs a commission-based affiliate scheme to extend its presence to many B2C sites. In terms of business development, however, it adopts a different strategy. Unlike Amazon.com, which has expanded horizontally to become a super e-store, 1-800-FLOWERS.COM has expanded vertically to sell various gift products. In other words, it has remained a specialized gift seller rather than a generalized super e-store. In addition to flowers and various gift products, 1-800-FLOWERS.COM also sells digital gift certificates. Upon receiving a prepaid digital gift certificate by e-mail, the recipient can use it to buy flowers or other gifts at 1-800-FLOWERS.COM.

Selling Digital Goods: EMusic

Obviously, digital goods (e.g., digital songs) are particularly suitable for e-commerce because they can be distributed directly through the Internet. As an example, let us study EMusic (http://www.emusic.com), which sells music (in MP3 format) through the Internet using a subscription-based business model. Note that MP3 is a widely used file compression format for music. Like other e-retailers, EMusic provides a search engine, a product directory, and help functions. The music can be searched for by artist, musical style, and so on. Furthermore, EMusic provides different charts to show the top songs and artists. Because of the special nature of digital goods, there are a number of distinctive features compared with other e-retailers that sell physical goods. In particular, subscribers are allowed to download as many songs as they want. Note that this feature is possible because, unlike physical goods, digital goods have extremely small variable costs. Upon downloading songs, a customer can also save them in compact discs or other devices for personal use. To download, listen, and organize MP3 songs, a software application called the audio player is required. In fact, many audio players are available free of charge on the Internet. Like other e-retailers, EMusic also uses an affiliate scheme to attract new customers or subscribers.

Selling Banking and Account Aggregation Services: Ascencia Bank and Yodlee

Besides selling tangible goods (e.g., books and flowers), the Web is also suitable for supporting different types of services (i.e., “intangible goods”) such as consumer banking services. In recent years, nearly all conventional banks have introduced supplementary Internet-based banking services. At the same time, a number of cyberbanks have been established that only serve their customers through the Internet (i.e., they do not have any physical branches). Nevertheless, they can still provide a wide range of banking services. Because they do not need to maintain any physical branches, significant operating costs can be saved, thus they can usually offer better interest rates than traditional banks. In fact, this is one of the key selling points of the cyberbanks. A U.S. cyberbank, Ascencia Bank (http://www.ascenciabank.com), serves as an example here. From the perspective of the services it provides, Ascencia Bank is almost the same as a conventional bank except that all banking services are provided via the Web or telephone. In particular, like other conventional banks, deposits up to a certain amount is protected (i.e., insured). Different types of bank accounts, such as personal and business accounts are available. Through a Web browser, users can check their account statements and transfer money at any time and anywhere. Of course, information access is protected by means of a username and password and other Internet security techniques (e.g., secure socket layer, discussed later). If a user has several bank accounts, he/she can access them through a single interface. Other services provided by Ascencia Bank include online investment (e.g., buying and selling stocks) and bill payment. In terms of help and communications, users can contact the bank by phone, e-mail, and fax.

A number of account aggregators, such as Yodlee (http://www.yodlee.com), have emerged in recent years (Deitel et al., 2001). These services allow users to access multiple bank accounts (even accounts from different banks) in an integrated manner via a single interface. In fact, Ascencia Bank also provides this aggregation service in cooperation with Yodlee. The basic operation is as shown in Figure 3. After authentication, the account aggregator serves as a proxy to get the information from the required banks for the user. The information is aggregated and possibly processed before being displayed to the user. Obviously, this service offers some advantages to consumers. It also presents a security issue, however, because users need to release sensitive personal information (e.g., account passwords) to the account aggregator for authetication purposes.

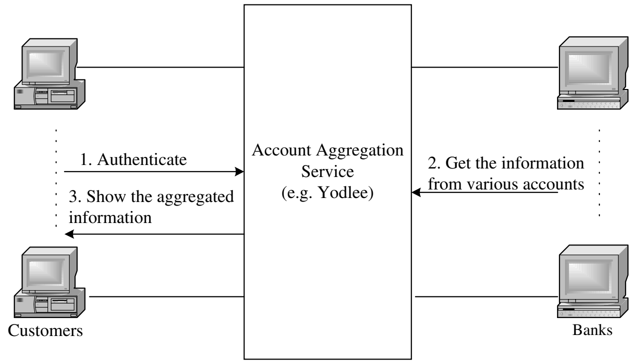

1. Authenticate
3. Show the aggregated information
Account Aggregation Service (e.g. Yodlee)
2. Get the information from various accounts
Customers
Banks

Figure 3: Account aggregation service.

C2B E-commerce

The traditional retailing model is seller-centric, which means that the consumers need to find out what each seller can offer. With the advent of the Internet, a buyer-centric model becomes feasible. In contrast to the seller-centric model, in the buyer-centric model each seller needs to find out what the consumers want. A representative example is Priceline.com (http://www.priceline.com), which provides a “name your price” service (or a demand collection service). For illustration purposes, the basic operation of the service is shown in Figure 4. Imagine that a traveler wants to book a five-star hotel room in a city. To use the service, he/she should provide the booking information and, most important, the desirable price. He/she also needs to provide a credit card number for payment. According to his/her requirements, Priceline.com searches for a hotel (supplier) that can satisfy his/her needs while fulfilling the required profit margin. Upon finding the supplier, Priceline.com gains the price difference as a service charge. If such a supplier cannot be found, a negative reply is returned to the consumer. The basic rule in using the service is that once a supplier is found, the buyer cannot refuse it (i.e., no objections and no refunds) and the credit card provided by the consumer will be charged. The products sold by Priceline.com have a special price–time relationship. Basically, the suppliers are operating with a high fixed cost but a relatively low variable cost. Furthermore, if a product cannot be sold before a certain deadline, the resource will be wasted (e.g., a vacant hotel room). For this reason, the seller may be

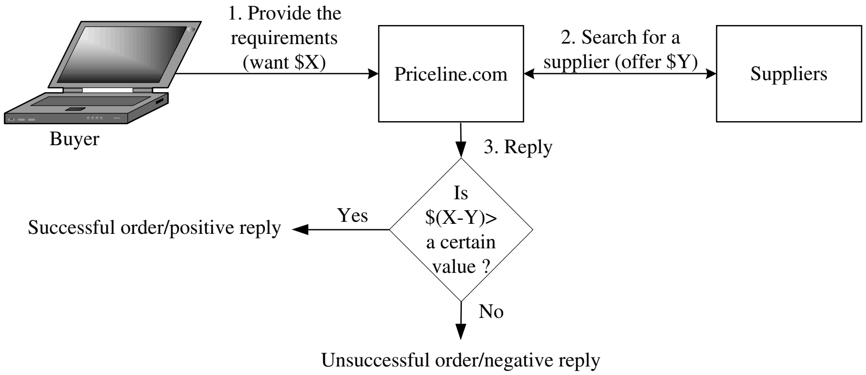

1. Provide the requirements (want $X)
Priceline.com
2. Search for a supplier (offer $Y)
Suppliers
Buyer
3. Reply
Is $(X-Y)> a certain value ?
Yes
Successful order/positive reply
No
Unsuccessful order/negative reply

Figure 4: The basic concept of Priceline.com's business model.

willing to accept an extremely low price—one much lower than the normal price but that is sufficient to cover the variable cost and perhaps part of the fixed cost. Apart from reserving hotel rooms, Priceline.com also sells other products including airline tickets and even home mortgages using a similar mechanism (i.e., the name-your-price approach).

C2C E-commerce

eBay (http://www.eBay.com) provides a Web-based auction service for consumers to trade goods, especially preowned goods, among themselves. Traditionally, classified advertisements are a common channel for selling used goods among consumers. Using Internet technologies, eBay offers an alternative and more effective solution. The basic operation is shown in Figure 5. To sell something, a seller initiates an auction by providing the required information. Buyers can search for the items they want and then bid for them. In most situations, the English auction method is used. This means that the buyers keep on bidding before the deadline; the buyer with the largest bid price gets the goods. For the English auction method, the bidding information is announced so that any interested buyer can easily check the latest bid and the bidding history. Based on the bidding information, a buyer can decide whether to submit a bid.

Besides product information, buyers can also check the track record of a seller, which is based on ratings provided by successful bidders having bought something from the seller. This feature is particularly useful, because in most cases buyers do not know the seller. In the simplest service, eBay only acts like a middleman to facilitate the consumer-to-consumer trading process. Upon completing an auction, the auction result is sent to the seller and the successful bidder by e-mail. Subsequently, the seller and the successful bidder need to make further arrangements to complete the deal (e.g., arrange delivery and payment). In terms of revenue, eBay gains a commission for each auction. As an additional service, eBay also provides an Internet-based payment method through PayPal (an associated company), which allows the seller and successful bidder to handle the payment on the Internet. Similar to many other auction sites, eBay also offers an escrow service. With this service, the successful bidder can pay the seller through eBay, but the seller will not get payment unless the goods have been delivered according to the agreement. Besides the English auction service, eBay also supports other auction methods (e.g., Dutch auction), and it provides specialized auction services (e.g., for selling motor cars) and live auction services.

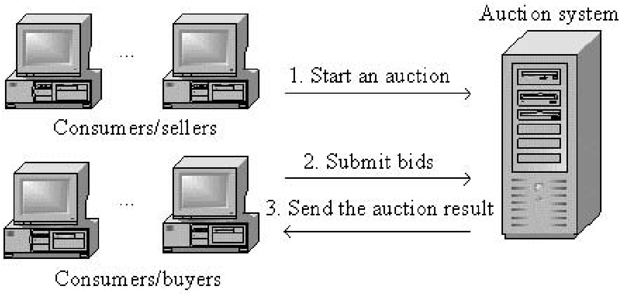


Figure 5: A consumer-to-consumer auction service.

In fact, eBay provides not only Web-based auction services but also a consumer-oriented community. At eBay, consumers can chat with each other in the online chat rooms and participate in many discussion groups covering various topics. All these community-oriented features aim at enhancing the stickiness of the Web site and strengthening customer loyalty. Apart from the auction services, eBay also has an associated company called Half.com, which allows consumers to sell used goods at a fixed (low) price instead of using an auction.

Supplementary Services

Other supplementary services exist to support consumer-oriented e-commerce. Two examples are e-marketing services (e.g., direct e-mail services) and e-catalog or Web hosting services. In the following sections, two case studies are presented.

E-catalog/Web Hosting Services: Bigstep

In general, small companies and home businesses may find it difficult to develop their own B2C Web sites from scratch. Fortunately, they can still enjoy the benefits of e-commerce through the use of Web hosting services provided by companies such as Bigstep (http://go.bigstep.com). In particular, these services help

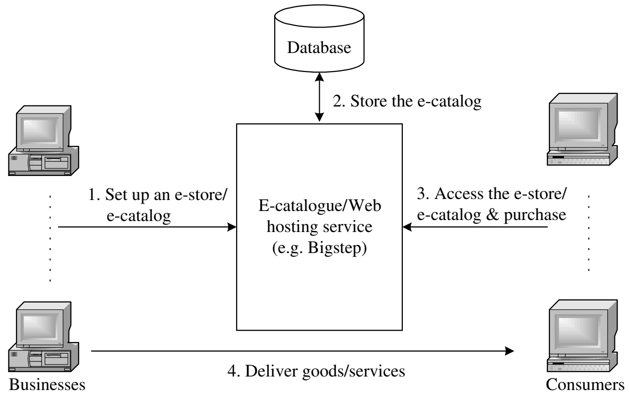

Database
2. Store the e-catalog
1. Set up an e-store/ e-catalog
E-catalogue/Web hosting service (e.g. Bigstep)
3. Access the e-store/ e-catalog & purchase
4. Deliver goods/services
Businesses
Consumers

Figure 6: Example of an e-catalog and Web hosting service.

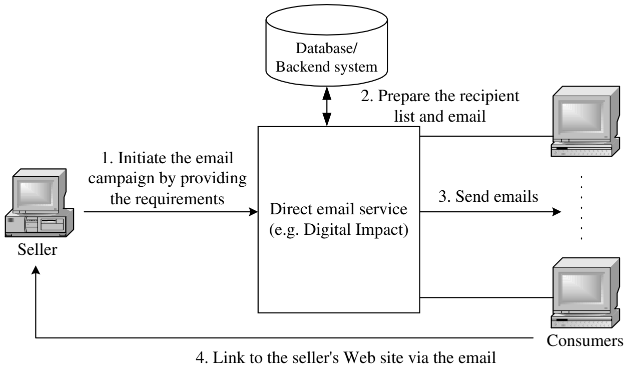

Database/ Backend system
2. Prepare the recipient list and email
1. Initiate the email campaign by providing the requirements
Seller
Direct email service (e.g. Digital Impact)
3. Send emails
Consumers
4. Link to the seller's Web site via the email

Figure 7: Example of a direct e-mail service.

small companies and home businesses to set up a Web site and an e-catalog cost-effectively by following standard procedures.

Figure 6 shows an example of the Bigstep service. By paying a monthly charge, a company can build a B2C Web site and then host it at Bigstep. Through the Web site located at Bigstep's server(s), the company can receive customer orders. By using the hosting service, companies can concentrate on their core businesses because their B2C sites are maintained by experts. To develop an e-catalog, a user can choose a store template together with other design parameters. It can be customized with descriptions and images, and so on. More sophisticated users can even add their own client-side programs (e.g., Hypertext Markup Language and JavaScript codes). Apart from providing a service for setting up a basic e-catalog, Bigstep also provides users with other e-retailer features, such as shopping carts, ordering handling facilities, and e-mail functions. In addition, users are provided with various reports to evaluate business performance and to understand the customer buying behavior. Bigstep also helps its users to market their e-stores by sending electronic newsletters to consumers and registering the sites with popular search engines. To a certain extent, Bigstep functions like a traditional shopping mall because it allows companies to set up independent e-stores at a centralized location. Note, however, that unlike a traditional shopping mall, e-stores can be set up by companies at different locations around the world.

Direct E-mail Services: Digital Impact

To become a successful e-commerce company, it is important to build a strong brand in cyberspace (Carpenter, 2000) and to use a wide range of marketing services to promote goods or services to the consumers. One of the effective marketing services is direct e-mail. There are three types of companies associated with direct e-mail services as described in Deitel et al. (2001). First, companies such as PostMasterDirect (http://www.postmasterdirect.com) keep recipient lists by encouraging consumers to register for inclusion in their database. Second, companies such as Digital Impact (http://www.digitalimpact.com) help businesses to implement direct e-mail programs. Third, supplementing text-based e-mails, some companies such as Avalon (http://www.mindarrow.com) provide multimedia attachments (e.g., animated demos and digital videos).

As an example, the service provided by Digital Impact is presented as follows. Direct e-mail is a focused (one-to-one) marketing technique). (See the white papers at http://www.digitalimpact.com for details, particularly Digital Impact, n.d., cited in the reference list.) Figure 7 shows the basic concept. A business first initiates an e-mail campaign by providing the marketing requirements. According to the requirements, Digital Impact prepares the promotional e-mail and the recipient list with the assistance of a database. The e-mail, possibly with multimedia attachment(s), is sent to the recipients. Interested recipients then visit the business's Web site through the link in the e-mail. It is hoped that some of the visitors will make a purchase. Moreover, the business can potentially identify each customer through the e-mail link and monitor his or her behavior. For example, the business may find that a particular customer is interested in some products. This information would allow the business to send customized e-mails to that person in the future. To enhance the effectiveness of a campaign, each e-mail usually provides an easy-to-use forwarding function to encourage the recipients to forward the e-mail to other people.

ENABLING TECHNOLOGIES FOR CONSUMER-ORIENTED E-COMMERCE

There are basically five technical building blocks for consumer-oriented e-commerce systems: communication infrastructure, programming languages, security techniques, payment methods, and other supporting technologies (Figure 8). An overview of these technologies

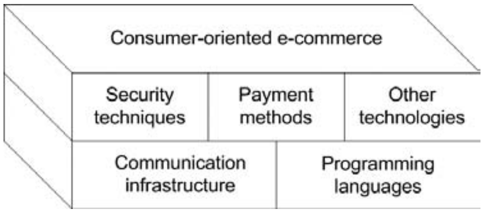


Figure 8: Technical building blocks for consumer-oriented e-commerce.

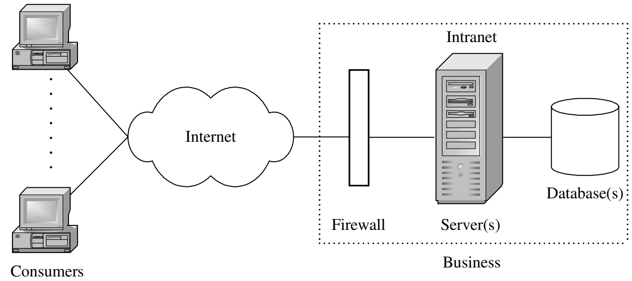


Figure 9: Basic system architecture for consumer-oriented e-commerce.

can be found in Chan et al. (2001). Currently, the World Wide Web provides the basic communication infrastructure or framework for supporting consumer-oriented e-commerce. As illustrated in Figure 9, a consumer accesses a Web server through his or her Web browser using the Hypertext Transfer Protocol (HTTP). The client interface is formatted based on the Hypertext Markup Language (HTML). On the server side, various programs (e.g., written in Active Server Page, Java Server Page, Java Servlet, etc.) are available to enable a server to interact with the clients, as well as with backend databases. Essentially, these server-side programs are used to generate dynamic Web pages in general and customized e-catalogs in particular. Note that in many cases, "cookies" (i.e., client-specific data or information transferred from the client-side to the server-side) are used for tracking sessions so that a server can send customized responses (e.g., customized e-catalogs) to clients.

Obviously, security is extremely essential in e-commerce. With the advent of various security technologies, secure transactions can now be conducted over the Internet. These security technologies include encryption techniques to safeguard data confidentiality, digital signature methods to ensure data integrity, and the digital certificate framework to support authentication (Stallings, 1999). Gathering together various security technologies, a number of protocols have been developed to address security requirements. In particular, the secure socket layer protocol is designed to protect the transfer of sensitive information (e.g., credit card information) between a client and a server. Firewalls are typically used to safeguard an Intranet against external security attacks (see Figure 9). For business-oriented e-commerce, Internet Protocol Security (IPSec) protocol is employed to set up virtual private networks (VPNs) for securing inter-business transactions.

Payment is an important function for supporting e-commerce. At present, credit cards are the most popular payment method for consumer-oriented e-commerce. With the support of major credit card companies, an application protocol called Secure Electronic Transaction (SET) has been developed for secure credit card payment over the Internet. Besides using credit cards, electronic cash protocols and electronic check systems are also available to emulate physical cash and check payments on the Internet, respectively. Apart from the traditional payment methods, a number of micropayment schemes are being developed to process low-value transactions, such as paying a few dollars to see a short video. Details of these payment solutions can be found in (O'Mahony, Peirce, & Tewari, 1997).

Last but not least, a number of supporting technologies are becoming increasingly important in e-commerce. For instance, software agents are used to facilitate searching and negotiation (see the next section). Many electronic customer relationship management systems make use of data mining and artificial intelligence techniques to make e-commerce more intelligent and user-friendly. The eXtensible Markup Language (XML) has been developed to facilitate information exchange, searching, and data presentation.

FUTURE DEVELOPMENTS OF CONSUMER-ORIENTED E-COMMERCE

Inspired by some of the ideas presented in Chan et al. (2001) and Hartman and Sifonis (2000), the development of consumer-oriented e-commerce can be divided into four stages (Figure 10). In the first stage, the Internet allowed global connectivity, thus forming a global entity and facilitating effective and efficient communications (e.g., using e-mails and File Transfer Protocol). In the second stage, the invention of the World Wide Web allowed efficient information retrieval over the global Internet. In this early stage of e-commerce, most companies only employed the Web for publicity purposes because purchase orders could not be taken securely. The third stage overcame this limitation by supporting secure transactions over the Internet. Furthermore, it also allowed the integration of Web systems with existing information systems and business processes. It is expected that the last stage will be a ubiquitous and fully automated e-commerce system that is created by integrating the existing e-commerce systems, mobile computing technologies, and software agents. Generally, the existing e-commerce systems allow consumers to search for information easily and make purchases conveniently. The last stage tackles two remaining problems: automation and mobility. The following paragraphs highlight the current developments in these two areas.

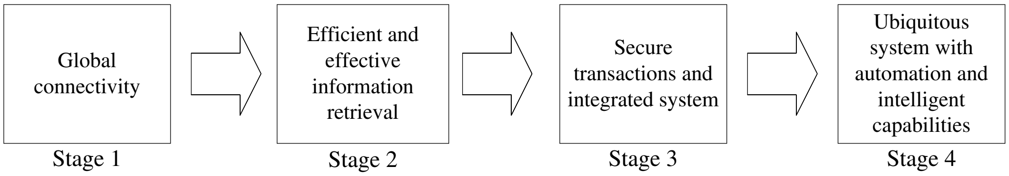

Global connectivity
Stage 1
Efficient and effective information retrieval
Stage 2
Secure transactions and integrated system
Stage 3
Ubiquitous system with automation and intelligent capabilities
Stage 4

Figure 10: Development of consumer-oriented e-commerce.

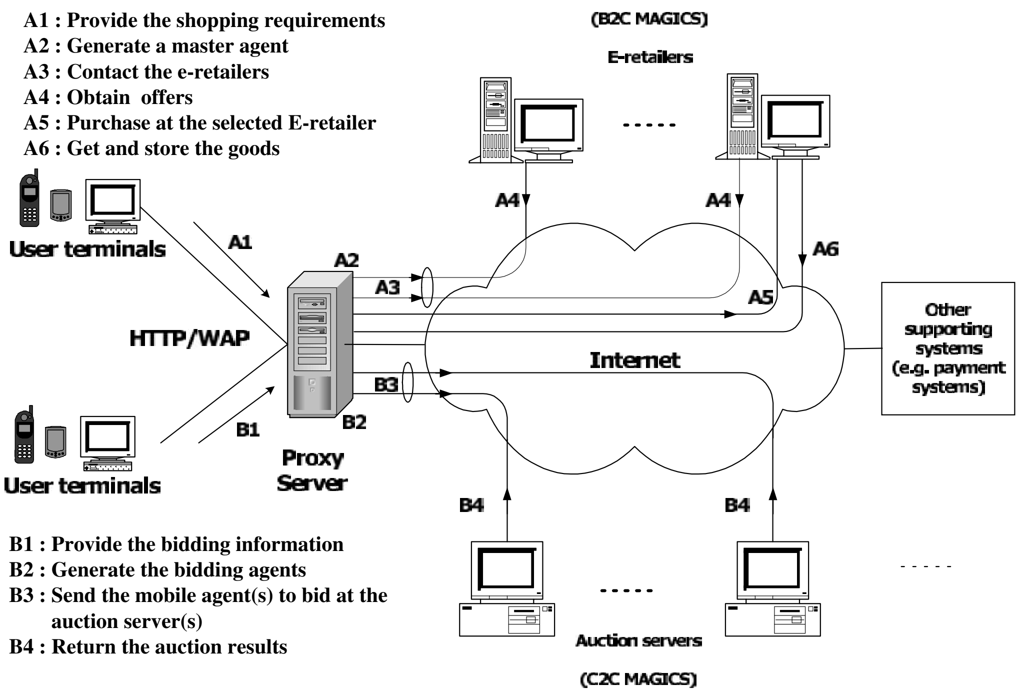

A1 : Provide the shopping requirements
A2 : Generate a master agent
A3 : Contact the e-retailers
A4 : Obtain offers
A5 : Purchase at the selected E-retailer
A6 : Get and store the goods
(B2C MAGICS)
E-retailers
User terminals
A1
A2
A3
A4
A4
A5
A6
HTTP/WAP
Internet
Other supporting systems (e.g. payment systems)
B3
B2
B1
User terminals
Proxy Server
B4
B4
B1 : Provide the bidding information
B2 : Generate the bidding agents
B3 : Send the mobile agent(s) to bid at the auction server(s)
B4 : Return the auction results
Auction servers
(C2C MAGICS)

Figure 11: Business-to-consumer and consumer-to-consumer MAGICS (Mobile AGent–based Internet Commerce System). (Adapted from Chan et al., 2002 and Ho & Chan, 2002.)

It is expected that software agents (or simply agents) will become important in next-generation consumer-oriented e-commerce systems (Moukas, Zacharia, Guttman, & Maes, 2000). Generally, agents are self-contained software programs that can perform specific tasks for human users. Because of their autonomy and communication capabilities, they are particularly suitable for assisting the consumer buying process. Early agent-based systems typically address the search and evaluation stages of the consumer buying process. In recent years, advanced agent-based systems have also been developed to handle more complex tasks, such as carrying out negotiations and even completing a sale transaction. Some examples include Tete-and-Tete and Kasbah (Moukas et al., 2000). In Kasbah, for instance, a seller or buyer can use a mathematical function to control how the selling or buying price should vary with time. Using the price–time functions, the selling and buying agents can negotiate with each other accordingly. There has also been considerable interest in investigating mobile agent-based systems (Lange & Oshima, 1998). Mobile agents can move around the Internet to complete the assigned tasks. Effectively, they can travel in cyberspace, like humans walking in a physical place. Three examples of mobile agent-based e-commerce systems are Nomad (Sandholm & Huai, 2000), MAgNET (Dasgupta, Narasimhan, Moser, & Melliar-Smith, 1999), and MAGICS (Chan, Chen, Dillon, Cao, & Lee, 2002).

Here we take MAGICS (Mobile AGent–based Internet Commerce System) as an example, which is being developed at The Hong Kong Polytechnic University. Inspired by various agent-based e-commerce systems, particularly Kasbah, Nomad, and MAgNET (please refer to the respective references), MAGICS currently provides two subsystems for B2C and C2C e-commerce. For the B2C MAGICS (see Figure 11), the buying process is summarized as follows based on the description in Chan et al. (2002). To use the service, a consumer connects to a special server through his/her Web browser or a mobile terminal (e.g., WAP phone). According to the buyer's requirement(s), some mobile agents are sent to look for the required goods and visit the e-retailers. The responses are then evaluated in accordance with the specified requirement(s), such as price. Finally, a buying agent is sent to the selected e-retailer to buy the goods. In addition to this shopping system, we are also working on C2C MAGICS, a mobile agent–based auction system (see Figure 11) (Ho & Chan, 2002). In this case, a consumer can create a mobile agent (or more than one), give it a bidding strategy, and then send the agent to the required auction server(s) to conduct the bidding. If several auction agents are generated for bidding at different servers, the agents may also coordinate the bidding themselves. In some advanced agent-based systems, it is even possible to create an intelligent agent that can learn about certain characteristics of its owner's behavior. This learning capability can be used to create many, more intelligent e-commerce services.

As mentioned earlier, another development in consumer-oriented e-commerce is to support mobility and ubiquitous access. Because mobile terminals and wireless networks have many limitations in respect of terminal capability, bandwidth, and so on, the existing Web system and protocol cannot be employed directly in the wireless environment (Mann, 2000). To build a mobile commerce system, it is necessary to address three fundamental issues. First, a high-speed access protocol is required for mobile terminals to utilize the wireless channel effectively and efficiently. Currently, mobile phone services are mostly supported by means of time division multiple access and frequency division multiple access. In the next generation mobile communication systems, a new technique called code division multiple access is likely to be used for supporting multimedia communications. Second, a simplified version of HTTP is required to allow mobile terminals to access a Web server through a proxy gateway. The Wireless Application Protocol (WAP) is devoted to this purpose. (See Mann [2000] for details). Third, a mini-version of HTML is needed to present Web or WAP pages on mobile terminals with tiny screen sizes. Currently, this is achieved by using the XML-based Wireless Markup Language. By employing these and other enabling technologies, various mobile commerce applications can be developed. Some examples include the following (Ahonen, 2002; Kalakota & Robinson, 2002; Turban et al., 2002):

Information inquiry—consumers obtain information to facilitate product comparison (e.g., inquire about product prices).

Location information—consumers get location information such as someone's location.

Mobile messaging—consumers read short messages through wireless terminals.

Mobile payment—a mobile terminal can function as a mobile digital wallet for payment purposes.

Mobile purchase—consumers carry out a sale transaction through pocket personal computers (e.g., buy a digital film ticket and even store the ticket in the terminal).

Multimedia entertainment—a typical example is the buying and playing of multimedia games through wireless terminals.

By combining mobile computing technologies and mobile agents, more innovative and useful consumer-oriented services can be provided. For instance, imagine that a mobile consumer needs to buy a film ticket as quickly as possible. He/she can create a mobile agent at his/her mobile terminal and then send it to the Internet to purchase the ticket in order to see the film at the closest cinema. Note that the buying process is fully automated because the agent can identify the consumer's location through a location tracking system and then search for the potential cinemas via other agents or search engines. (See Moukas et al., 2000, for similar examples.)

CONCLUSION

This chapter has provided an overview of e-commerce in general and consumer-oriented e-commerce in particular. With the aim of selling different goods and services on the Internet and facilitating the consumer buying process, many consumer-oriented e-commerce models have been developed. These e-commerce applications can provide many benefits to both consumers and businesses. To implement various consumer-oriented e-commerce applications, a client-server architecture is used, and the technical building blocks include communication infrastructure, programming languages, security techniques, payment methods, and other technologies. It is predicted that the current e-commerce system will evolve to a ubiquitous system that offers more automation and increasingly intelligent functions.

ACKNOWLEDGMENTS AND NOTE

This chapter was supported in part by a Competitive Earmarked Research Grant from the Research Grants Council, the Hong Kong Special Administrative Region, China under account number PolyU 5078/01E and by the Department of Computing of the Hong Kong Polytechnic University under account number 4-Z006. We thank Ms. Irene Ho, Mr. Peter Meldrum, and Ms. Perry Lam for their help in the preparation of this chapter. We also thank the reviewers for their valuable comments. Note that the trademarks and service marks mentioned in this chapter belong to their respective companies or owners. They are only used for identification purposes.

GLOSSARY

Account aggregators A system that allows a user to handle several accounts (e.g., various bank accounts) via a single (and possibly customized) interface.

Affiliate scheme A commission-based marketing technique for a business to refer its visitors or customers to another business over the Internet.

Auction A dynamic pricing technique for facilitating a sale transaction.

Business-to-consumer (B2C) e-commerce One type of e-commerce involving the sale of goods or services from a business to a consumer.

Consumer-to-consumer (C2C) e-commerce One type of e-commerce involving the sale or purchase of goods among consumers.

Consumer-to-business (C2B) e-commerce A buyer-oriented e-commerce model enabling a business to provide a customized solution to satisfy a consumer need.

Cookies Client-specific data or information forwarded from a Web client to a Web server (e.g., for generating a customized response).

E-commerce Sale and purchase of goods or services by electronic mechanisms in general or over the Internet in particular.

Extensible Markup Language (XML) A general markup language for users to specify their own tags.

Hypertext Markup Language (HTML) A markup language for formatting or showing Web pages by employing standard tags.

Hypertext Transfer Protocol (HTTP) A request/response protocol for supporting the Web system.

Internet Protocol Security (IPSec) A security mechanism for the Internet protocol (IP) to support data confidentiality and/or data integrity.

Mobile agent A mobile software program which moves around the Internet to complete a predefined mission automatically.

Micropayment methods Payment methods for processing low-value transactions.

Secure Electronic Transaction (SET) An application protocol for secure credit card payments over the Internet.

Secure Socket Layer Protocol A transport layer security protocol typically for protecting the connection between a Web server and a Web client.

Software agent A special computer (software) program capable of completing a mission in an autonomous manner.

Wireless Application Protocol (WAP) An application protocol for a mobile/wireless terminal to access a Web server through a gateway.

Wireless Markup Language (WML) A simplified version of HTML for showing Web (WAP) pages on mobile or wireless terminals (e.g., mobile phones).

CROSS REFERENCES

See *Business Plans for E-commerce Projects; Business-to-Business (B2B) Electronic Commerce; Business-to-Business (B2B) Internet Business Models; Business-to-Consumer (B2C) Internet Business Models; Click-and-Brick Electronic Commerce; Collaborative Commerce (C-commerce); Electronic Commerce and Electronic Business; Electronic Data Interchange; Electronic Payment; E-marketplaces; Global Issues; Online Auctions; Web Services.*

REFERENCES

Ahonen, T. T. (2002). *m-profits: Making money from 3G services.* Chichester, UK: Wiley.

Carpenter, P. (2000). *eBrands: Building an Internet business at breakneck speed.* Boston: Harvard Business School Press.

Chan, H., Chen, H., Dillon, T., Cao, J., & Lee, R. (2002, October). A mobile agent-based system for consumer-oriented e-commerce. Paper presented at the 4th International Conference on Electronic Commerce, Hong Kong.

Chan, H., Lee, R., Dillon, T., & Chang, E. (2001). *E-commerce: Fundamentals and applications.* Chichester, UK: Wiley.

Dasgupta, P., Narasimhan, N., Moser, L. E., & Melliar-Smith, P. M. (1999). MAgNET: Mobile agents for networked electronic trading. *IEEE Transactions on Knowledge and Data Engineering, 11,* 509–525.

Deitel, H. M., Deitel, P. J., & Steinbuhler, K. (2001). *e-Business & e-commerce for managers.* Upper Saddle River, NJ: Prentice Hall.

Derfler, F. J., & the editors of *PC Magazine* (2001). *E-business essentials.* Indianapolis, IN: Que.

Digital Impact (n.d.). *The personal touch: Making one-to-one e-mail marketing work for you.* Retrieved October 2002 from http://www.digitalimpact.com/docs/one-one_whitepaper.pdf

Etzel, M. J., Walker, B. J., & Stanton, W. J. (2001). *Marketing* (12th ed.). Boston, MA: McGraw-Hill/Irwin.

Hartman, A., & Sifonis, J., with Kador, J. (2000). *Net ready: Strategies for success in the e-conomy.* New York: McGraw-Hill.

Ho, S. K., & Chan, H. C. B. (2002, October). On the design of auction agents with different bidding strategies. Paper presented at the 4th International Conference on Electronic Commerce, Hong Kong.

Kalakota, R., & Robinson, M. (2002). *M-business: The race to mobility.* New York: McGraw-Hill.

Kalakota, R., & Whinston, A. B. (1997). *Electronic commerce: A manager's guide.* Reading, MA: Addison-Wesley.

Korper, S., & Ellis, J. (2001). *The e-commerce book: Building the e-empire* (2nd ed.). San Diego, CA: Academic Press.

Lange, D. B., & Oshima, M. (1998). *Programming and deploying Java mobile agents with aglets.* Reading, MA: Addison-Wesley.

Mann, S. (2000). *Programming applications with the Wireless Application Protocol.* New York: Wiley.

Moukas, A., Zacharia, G., Guttman R., & Maes, P. (2000). Agent-mediated electronic commerce: An MIT media laboratory perspective. *The International Journal of Electronic Commerce, 4*(3), 5–21.

O'Mahony, D., Peirce, M., & Tewari, H. (1997). *Electronic payment systems.* Boston, MA: Artech House.

Rayport, J. F., & Sviokla, J. J. (1994). Managing in the marketspace. *Harvard Business Review, 72,* 141–150.

Rayport, J. F., & Sviokla, J. J. (1995). Exploiting the virtual value chain. *Harvard Business Review, 73,* 75–85.

Sandholm, T., & Huai, Q. (2000). Nomad: Mobile agent system for an Internet-based auction house. *IEEE Internet Computing, 4,* 80–86.

Schneider, G. P., & Perry, J. T. (2001). *Electronic Commerce* (2nd ed.). Boston, MA: Course Technology.

Stallings, W. (1999). *Cryptography and network security.* Upper Saddle River, NJ: Prentice Hall.

Turban, E., King, D., Lee, J., Warkentin, M., & Chung, H. M. (2002). *Electronic commerce 2002—a managerial perspective.* Upper Saddle River, NJ: Prentice Hall.

Zwass, V. (n.d.). *Structure and macro-level impacts of electronic commerce: From technological infrastructure to electronic marketplaces.* Retrieved May 2002 from http://www.mhhe.com/business/mis/zwass/ecpaper.html

Convergence of Data, Sound, and Video

Gary J. Krug, *Eastern Washington University*

INTRODUCTION

In recent decades, video, data, and audio material have become increasingly combined into new forms that are more than the sum of their parts. This convergence of different forms of representation is made possible by many technological and social factors including the emergence of new technologies, a deregulated telecommunications industry, and a growing consumer, corporate, and institutional interest in multimedia. Many universities, museums, and corporations now include multimedia (including three-dimensional effects, animations, sound, and video) on their Web sites, and multimedia compact discs (CDs) are increasingly distributed either in place of or in conjunction with printed promotional materials. With increasing sales of digital video and still cameras, flatbed scanners, and inexpensive video, imaging, and audio editing software, multimedia production and deployment are becoming commonplace for many computer users. There remain limitations, however, to the development and spread of multimedia, and the social effects of multimedia use and deployment remain largely unknown.

Only with the advent of CDs, relatively high-speed dial-up connections, and broadband did the merger of data, sound, and video become possible on the personal computer, thus providing the most ubiquitous form of multimedia. This is changing, however, for multimedia has begun to appear in wireless forms as well. Increasingly, cellular phone systems are converging with graphic uses such as text messaging and Internet access. Both new digital phones and the hybrid personal digital assistants (PDAs) incorporate aspects of the personal computer with telephone technology. Corporate collaborations such as that of Verizon Wireless and Microsoft to provide wireless data services to telephone customers will further the development of this convergence, although it is unlikely that wireless will in the short term displace other forms of multimedia delivery. Nonetheless, such delivery systems may make these forms omnipresent in everyday life, delivering e-mail and Web sites as well as advertising wherever one is. A more likely outcome is that wireless services will function as supplementary forms to other multimedia uses extending the use multimedia into new contexts. The current deployment of 3G (third-generation) wireless telephony is driven largely by its ability to provide these functions and services.

Convergence

The convergence of data, sound, and video brings together different forms of information into an integrated, multimedia construction. The advantage of providing these three dimensions of sensory experience at one time is that a depth of information about a topic may be experienced in a single page or demonstration. These additional forms of representation may convey more information about the writer's thoughts or attitudes, or they may provide information not easily described or related in writing. For example, some limitations of writing, and hence of purely textual information, have historically been the lack of clues about authorial intention and the emotional inflection of the words and phrases that appear in the text. These additional dimensions of the communicative act must be inferred through contextual clues. An equivalent of this contextual information may be given in video or audio forms or through other frameworks that create a relationship between the reader and the text so that particular understandings of the text are more likely than others. Such information stands in addition to traditional textual framings that also contribute to the way in which a text is read and understood by readers. These traditional framings, such as a knowledge of the author, an expectation of the genre, and previous experience with a publisher, to name a few, give the reader clues about whether a particular work is fiction, science, news, and so on and thus influence the range of possible meanings that the reader may construe.

Multimedia information may thus facilitate some communication by allowing for the inclusion of additional contexts for communication. Some kinds of complex information are best conveyed in images and illustrations, and the accompaniment of video within a text delivers a high density of information that may elaborate the intended message of the written text. Whereas print has long been able to produce illustrations, pictures, and diagrams to accompany textual material, early digital media did not. The subsequent accompaniment of text with sound and video in addition to images requires a reexamination of the relationships between various media as well as a reconsideration of what a text is and how it functions within larger semiotic systems.

Multimedia formats for creating electronic pages, especially those deriving from HTML (hypertext markup language), are often referred to as being "interactive" media because the viewer has some choice in controlling what parts of the multimedia are presented and the order in which this presentation occurs. The word interactive is a misnomer, however, implying that one has a degree of control over the content of the message, that is, that the message is in fact changed or created by the viewer. Meaning is not so indeterminate in most multimedia presentations. As noted earlier, various other technologies within the text, as well as the reader's experience with prior texts, overdetermine the general frame in which the text is read and thus guide the reader to a preferred meaning of the text. In many settings, the "interactive" aspect of media is completely subsumed by the form of the text itself: Short, clear, and immediately conveying the intended message, the multimedia page often discourages ambiguity and multiple interpretations.

MULTIMEDIA IN THE CONTEXT OF REPRESENTATION

Multimedias contribute to the creation of virtual worlds, that is, representations of real or imaginary places complete with images, video, and sound. In this way, they may present varying degrees of verisimilitude of the real. One may thus have a virtual visit to another place or time. This may be helpful for people considering a move to a city, company, or school that they have never visited. When using multimedia for these sorts of virtual visits, however, one should remember that the materials presented are selective and most certainly are designed to present the best possible face. That is, all productions are engineered to produce a specific effect and to guide the viewer or reader to a specific subjective experience of the text and so to some specific meaning, attitude, or relationship to the presentation.

The history of media in general is also the history of divergences and convergences of different modes and formats of representation. Painting and music existed in social and cultural contexts that guided the person experiencing them toward particular kinds of meaning: For example, a painting in a museum is a work of art, whereas a painting on a wall is an act of vandalism. With telecommunications, the contexts of communicative forms are often missing from the message. Multimedia create new contexts through associating aspects of the message with other media forms. A work of music accompanying a text will alter how the meanings of the text are made. Almost from its earliest days, film has exploited the semiotic relationship between sound, music, image, and narrative to achieve specific ends.

The elements of a multimedia construction together constitute something new that is more than the sum of its parts. Text, music, images, and video create a field of representation in which each combines with the others to create a larger impression that in turn adheres to the topic of the multimedia presentation and colors the entire experience. As such, a study of multimedia should include an examination of earlier attempts to integrate sound, visual effects, and even live action, as in opera, ballet, and theater. In appearing with text, however, the multimedia construction borrows as well from some conventions of reading and writing while introducing aesthetic traditions and grammars of sense-making from the video, filmic, and photographic conventions. The relationship between multimedia and other forms extends beyond simple layout and page design and already is creating new forms for each of these constituent elements as well as new experiences of reading and creating multimedia texts.

All forms of representation—writing, pictures, diagrams, video—are limited in what they can convey and in the sorts of messages that can be inferred from them. Furthermore, the function of some representations is not to communicate some specific information but to serve what James W. Carey (1991) called a ritual function, that is, providing some part of the mundane background of shared everyday life. This view of communication and media is shared by Luhman (2000). Beyond the phatic dimension of communication, ideological functions contribute to the formation of subjectivities of identity by presenting templates for normative forms of class, race, age group, gender, and so on. As multimedia grows in popularity and as large corporation increase the use of multimedia in cross promotion of products and services, a greater conventionality of presentation will emerge, and these conventions will become associated with the representations of corporate products. Thus, the whole of the primary aesthetic of multimedia may well become standardized in the forms of advertising and public relations.

TECHNOLOGIES

In assessing the growth of multimedia, it is important to understand that technologies do not in themselves create new fields and activities. Rather, technologies emerge into specific forms based on the culture and time in which they are developed. Much of the computer revolution is an offshoot of research and development in military applications, and this applies as well to technologies associated with multimedia. This historical and cultural grounding of media suggest as well that the uses to which new technologies are put are also influenced by the social and cultural settings into which they emerge. The relationship between technology and the social is not determined by either but arises from the relationship they form with one another.

Jacques Ellul noted that "Each new machine disturbs the equilibrium of production; the restoration of equilibrium entails the production of one or more additional machines in other areas of operation" (1964, p. 112). Equilibrium is, however, never established. Technology leads to the endless displacement of existing practices in the constant introduction of new devices and their use, but, as with the promise of the "paperless office," the expected goal is seldom achieved. For example, the introduction of the computer unleashed unprecedented amounts of information into the social world. This necessitated the establishment of systems for exchanging information (the Internet, local area networks, etc.). The proliferation of information managers and workers in turn created a need for simpler, more user-friendly interfaces, and so on. Each material development of a technology creates the need for subsequent machines, devices, and technologies, and each of these in turn influences changes in the manner and form of the workplace, the relations of workers to the job, to each other, to the institution, and so on. Technological change thus appears as an ongoing renewal and recreation of the world, but technology also comes to appear as an autonomous force in the world.

The following history briefly touches on some of the key components and moments in the development of digital multimedia.

History: Phones, Photographs, Video, and Television

Data in the form of digital telegraphic signals was the first electronically transmitted information. Patented in 1839, the telegraph established the pattern for subsequent communication system deployments, quickly becoming important in domestic and international politics and business in the industrialized nations. Before the American Civil War, there were already attempts at laying transatlantic submarine cables, and by 1872 the British had successfully linked their Australian holdings to England with telegraph lines. Major international rivalries developed over telegraph monopolies, particularly between England and Germany in the late 19th century. At the time, there was considerable expectation that these electronic linkages of people as well as other advances in communications, such as the Penny Post (1840), would unify the world and usher in a new age of peace and universal understanding.

Neither the amount of communication nor the speed at which it took place could produce the shared values and understandings necessary for a utopian world, but communications did establish linkages between economic markets. Among the great successful commercial applications of telegraphy was the ticker tape that conveyed stock prices from major markets to subscribers at speeds approaching real time transmission. While this technology centralized and linked markets together, it also made possible sudden shifts in the market by speeding up trades, and the increased speed may have contributed to the emergence of trading in futures.

Despite its popularity with business and governments, telegraphy remained a technology with which most people would have little direct contact. The telephone would be the first technological system to reach into the homes and lives of ordinary people, and although the basic device was demonstrated by the middle of the 19th century, it did not become commonplace until the 1900s. Telephone technology developed rapidly, creating the physical infrastructure of main lines, switchboards, and drops to individual homes and offices. By the 1930s, rotary dial telephones were introduced making automatic mechanical switching of telephone calls possible and beginning the automation of the system, eliminating the manual connection of calls by operators.

Photography emerged in 1839 with the metal plate process developed by Niepce and Daguerre in France and the paper negative calotype developed by Talbot in England. Bayard, another Frenchman, also developed a paper negative system at the same time. Although high-quality permanent images could be made by both processes, no inexpensive means of producing large numbers of copies existed until the 1870s. As such, photography was limited in its uses and dissemination. Although illustrations had accompanied text from at least late antiquity or early medieval times and woodblock printing was widely used, photographic reproduction did not commonly appear in books until the late 19th century. Among the early publications were Charles Darwin and O. G. Rejlander's *The Expression of Emotions in Man and Animals* (1872) and John Thomson's *Street Life in London* (1877). These works, and other uses in fields such as astronomy and biology, demonstrated that photographic images and written words in combination could convey scientific observation as well as social commentary in more powerful forms than had previously existed.

The use of the apparently "genuine" image created a new sense of verisimilitude of pictures: A photo does not lie. Over time, however, the proliferation of images as well as photographic effects (for example, multiple exposures) undermined the credibility of the medium itself. Nonetheless, photography exerted powerful social effects. Darwin's book marked the beginning of a tradition of presenting the empirical evidence of science in images which the text expanded and commented on. Thomson's works stood at the beginning of a tradition that would develop into the social reform movements and muckraking journalism and so influence social development by conveying in realist detail the lives of the poor. Additionally, photography made affordable for many people the individual and the family portrait.

Moving images were first developed in the 1890s and enjoyed immediate popularity worldwide. The motion picture extended the illusion of verisimilitude across time: Movies portray not a frozen, fixed moment but a duration or span of time. Sound was synchronized with the images in the 1920s, becoming widely used in the 1930s and 1940s, and color film began to be widely used during World War II. Still dependent on film stock and projectors, however, moving pictures were largely confined to the social spaces of cinemas. Home movies in the popular 8-mm format were limited in resolution and short in duration, and most home film equipment was silent.

Although some advances were made in the 1920s and 1930s, it was not until after World War II that electronic transmission of sound combined with video became viable. Philo Farnsworth and Vladimir Zworykin demonstrated the technology for capturing and transmitting images electronically in the 1920s, and in 1939 public broadcasting began in the United States. Interrupted by the war, commercial television broadcasting on a large scale began in the in 1947.

Television began the domestication of video, bringing it out of the theater and into the home. Television, itself incorporating several existing technologies, would prove to be one site on which further technologies would converge. One of the most significant was the development of the VCR. First released commercially by Ampex in 1956, the video tape recorder (VTR) underwent several refinements on its way to becoming the video cassette recorder. Although Sony had experimented with home use VTRs, Philips introduced the first VCR for home use in 1972, and by 1977 the VHS system was marketed. In the following decade, the VCR exploded in popularity, and for the first time, television could be recorded and edited. The VCR introduced wholly new practices to video watching such as time shifting, that is, watching broadcasts at later times. This practice enabled another: recording one channel while watching something else. To the dismay of advertisers, this technology also allowed people to engage in "commercial zapping," that is, fast-forwarding through commercials.

Capturing two or more channels simultaneously became more important as cable television became a major mode of distribution increasing the number of available channels and fostering the growth of niche market programming. Although CATV (community antenna television) had been in use since the late 1940s, it was not until the 1970s that satellite links to cable delivery became widespread. This new system not only greatly increased the number of channels available but also produced the infrastructure for broadband terrestrial delivery of television and other services. Video on demand (VOD) and other forms of pay-per-view systems restructured the television industry, and, as television distribution became increasingly digital, other convergences of technology would produce both new industry practices and new ways of using the devices. In the 1990s, the television and the personal computer (PC) began to appear to be converging into one device.

Socially interactive systems, such as Qube TV, did not become popular for various reasons and have largely disappeared. Interactive television technology is now largely limited to the choice of product for consumption (such as pay-per-view) and the limited provision of textual information about a program. To date, television and PCs remain quite different devices, although the emergence of digital television fostered the rise of devices such as TIVO, which are capable of capturing and storing digital video for later playback to a television or for capture on a computer.

Television and its associated technologies did, however, introduce the moving image into everyday life and produced a level of understanding of both the grammar of the video form and the technology associated with it. Rather than being an occasional experience at the cinema, video became a significant part of everyday life as well as a major source of revenue for the motion picture industry. The Video Software Dealers Association reported that the videotape rental industry accounted for $7 billion in rental revenue in 2002. In comparison, DVD (digital versatile disk) rentals accounted for $1.4 billion in the same year.

Hardware Development

As early as the 1970s, researchers such as Douglas Englebart were exploring ways of using shared-screen telephone conferencing in conjunction with other computer technologies. Envisioned primarily as a tool for the workplace, these early forays into multimedia laid the groundwork for many subsequent developments in computing, and they contributed significantly toward the idea of making computers accessible and user-friendly. Many scholars have recognized the contribution of Vaneveer Bush and particularly of a piece Bush published in the *Atlantic Monthly* in 1945 entitled "As We May Think." In this work, Bush envisioned an integrated workstation which, using predigital technologies, would allow researchers to access vast stores of information and then store the cross-references, which he called "associations." Thus, a conceptual form of hypertext emerged while computers were still in their infancy.

Personal Computers

The growing use of increasingly powerful and versatile PCs capable of high-resolution image production combined with sound was driven, in part, by the popular videogame industry. Growing sophistication in software, sound, and graphics, quickly overwhelmed the limited storage capacity of 1.44-MB floppy disks and fueled the growing use of alternative storage media such as the CD and DVD.

The inclusion of images, sound, and video with textual materials did not become viable until various hardware and software criteria were met. Images and video are processor and memory intensive. Digitized, or sampled, music of high-quality creates large files. Thus, in the SOHO (small office, home office) and consumer markets, multimedia helped generate the demand for powerful processors, extensive video and system RAM, and large storage systems, especially hard disk drives (HDDs) and optical storage systems such as CDs and DVDs. In processors, the MM technology from Intel and the 3DNow technology from AMD incorporated special instructions into the processor's language enabling faster handling of image processing and video compression and /decompression. Apple Computer was also a major player in the development of image and sound processing equipment. As early as 1984, using the Motorola 68000 processor chip, Apple was establishing itself as a preferred platform for many video and graphics creators and users.

CDs and DVDs

The CD is a transportable, high-capacity storage system, holding about 640 MB of data on a standard disk. Developed by the Phillips Company in 1979 and initially released as a videodisk system, the audio CD was launched in 1983. The DVD is capable of holding up to 4.7 GB of data on each side of a disk. This means that about 133 minutes of high-quality video and sound or up to eight hours of music can be stored on a DVD. Advances in DVD technology are set to increase this to 17 GB or higher.

The CD remains a common multimedia delivery format, although increases in bandwidth continue to erode its popularity and shift delivery away from this last significant link to the physical medium. The CD format provided a significant improvement in transportable storage systems, enabling for the first time the distribution of large amounts of text and multimedia in a small and easy to use system. For the first time, traditional publications such as encyclopedias were no longer limited to text-based forms. Examples of music, high-quality images, animations, and fragments of historically significant video are now standard components of the multimedia encyclopedias. As such, the encyclopedia is becoming indistinguishable in form and in the ways that it is used from any other multimedia product.

Bandwidth

The limitation to large size, high-speed data transfer has traditionally been the scarcity of bandwidth in existing communications networks. In the early days of data transfer, 300 baud was satisfactory for simple exchanges. Images, however, even if compressed with "lossey" algorithms such as JPEG files, require large amounts of bandwidth relative to text. Files such as bitmaps, TIFF, and GIF files are even larger and require even more bandwidth to transmit. Current video transmission rates on most Internet providers are geared to what may effectively be conveyed over a 32-KB system. As a result, the frame rate and resolution are reduced, causing significant losses in smoothness and quality of image. As of 2000, about 80% of U.S. homes did not have broadband service, and little progress in rollout has been made in the subsequent two years.

The limitation of bandwidth was addressed by various multiplexing systems, beginning with frequency-division methods that gave way to time-division technologies. These significantly reduced the size of transmissions and enabled larger numbers of users to simultaneously share the same physical lines.

The development of a high-capacity backbone may be traced to the emergence of the L1 coaxial backbone in 1946. Coaxial cables greatly increased the carrying capabilities of telephone lines over copper pairs. As digital technologies multiplied in the 1980s and 1990s, a dedicated digital backbone for both voice and data was deployed using a combination of digital multiplexing over terrestrial lines, satellites, and microwave relay towers. These are now supplemented with fiber-optic cables. In 2000, global Internet capacity surpassed voice call capacity worldwide for the first time, demonstrating that data transmission requires design changes to the telecommunications infrastructure. One such change has been the development of MPLS (multiprotocol label switching) to allow the more efficient routing of data packets within Internet routing computers. This technology allows the sender of a packet to specify the priority to be given to a packet based on the needed QoS (quality of service) and may even specify some steps in routing the packet to improve speed. For example, data transmission may suffer pauses and slowdowns without the user noticing, but VoIP (voice over Internet protocols) and video would suffer noticeable gaps and unacceptable interruptions. Thus, with transmission of these sorts of data, the QoS demand will be higher, and the data packets will receive a higher priority in routing at each node of the transmission.

Further transmission technologies more suited to small, localized settings such as homes and small business continue to develop, examples being various DSL (digital subscriber line) technologies: ADSL (asynchronous digital subscriber line), VDSL (very high data rate digital subscriber line), and so on. New developments also offer the possibility of high bandwidth transmission through existing power lines and circuits.

Recent extensions of the fiber optic network such as the Southern Cross extension in Australasia and the FLAG (fiber-optic link around the globe) submarine cables have significantly increased data bandwidths and allowed the replacement of geosynchronous satellite relays, which produce a time lag in transmission, with faster, more reliable landlines. In addition to providing significant increases in bandwidth over satellite transmission, terrestrial cables are less affected by atmospheric and solar conditions and have fewer security problems associated with data interception.

The Last Mile

The great bottleneck in providing high-speed data services has always been what is called "the last mile" of the network. However advanced the data transmission backbone may be, the linkage between that source of high-density data transfer and the end user often presents problems. In many settings, population density is not high enough to justify extending the high-speed data connection directly to the doors of individual consumers, especially domestic consumers. Direct satellite connection is often a viable solution in these cases. Indonesia, which is an archipelago of many islands, was one of the first countries to exploit satellite technology as a telecommunication backbone. In other countries, a high population density combined with national interests in developing extensive digital communications technologies have engendered plans to upgrade delivery significantly. Singapore, for example, is rolling out fiber-optic cabling to nearly every home and office in the country and has achieved some of the highest rates of computer ownership and Internet use in the world.

Bandwidth remains an issue in providing data services over existing phone lines. Among the current systems for high bandwidth delivery within the U.S. are ISDN (integrated services digital network), DSL, and T1 lines, which use the existing telephone system, digital cable, and direct satellite. ISDN and T1 both require dedicated lines, whereas DSL uses a frequency division system of transmitting over the twisted copper wire pairs of a POTS (plain old telephone system). Although the speed of ADSL is related to the distance from the central switching unit, rates of data transfer in excess of one MB/s are possible. Other possibilities for the local loop include cable service which can offer digital speeds comparable to DSL. Although widely available, one drawback of cable modem service is that the bandwidth available to the individual user drops with increasing numbers of subscribers. Technologies are constantly being deployed, such as high speed data over ac power lines, which became commercially available in Scotland in 2003.

Within the United States, the loop between the individual user and the telephone switching office had been preserved by federal law and policy as a monopoly for local telephone companies. The Telecommunications Act of 1996 (Benton Foundation, 1996) deregulated this and other areas of the telecommunications market, allowing other businesses and corporations to compete in service delivery. Although technically allowing cable companies and even local utilities to provide telephonic systems to the home, to date few nontelephone companies have launched these services. The most successful nontelephone broadband service in the United States continues to be cable television. Most of Europe and many cities in the United States have established complete connection of digital telephone exchanges to local telephone service, however.

Overall, broadband has not been widely taken up in the United States. An estimated 70% of U.S. homes have broadband service available to them, but only 10 to 12% currently use such a service. Although there are an estimated 18.5 million DSL users worldwide, making DSL the most commonly used form of broadband access from home, this is a small fraction of the roughly 500 million personal computers estimated by the International Telecommunications Union to be currently in use (ITU, 2003).

Data Compression Technologies

One answer to the bandwidth problem has been the development of more efficient compression algorithms and, in some cases, hardware support for these. The Motion Picture Experts Group (MPEG) develops standards in video compression that also have applications to audio. Their standards are designated as MPEG and employ various algorithms and technologies for reducing the size of video and sound files. MPEG-1 is generally used in transmissions up to 1.5 MBS. MPEG-2 is the standard compression technology of DVDs and digital television. MPEG-3, also commonly known as MP3, enjoys great popularity as a music compression format.

According to some industry analysts, a predicted growth in broadband home access combined with new data compression technologies will enable a much larger deployment of video on demand over the Internet. Traditional media companies such as Disney, AOL Time Warner, and Viacom will look to Internet delivery of their programming to increase market share. At the same time, other companies will be looking to expand into delivering Internet services over televisions. Clearly the emerging convergence of traditional broadcasting and the Internet services will be speeded by the change from analog to digital television, and continuing convergences of technologies, corporations, and content will bring PCs and television closer together in the near future.

USES

In general, the convergence of technologies emerges in the context of changing alignments of social institutions. Lines between traditional corporations blur, as do the divisions between previously distinct genre and forms of information. The deregulated environment in the United States following the 1996 Telecommunications Act significantly increased the number of participants providing services and allowed companies to develop innovative combinations of services.

Rather than producing the diversification of service providers that was predicted by supporters of deregulation, however, the industry has seen a growth in consolidation and centralization of ownership through a series of large mergers, the most famous of which was the AOL–Time Warner Merger in 2001. This deal surpassed the previous record for communication corporation mergers set in 1999 by the Viacom takeover of CBS. Media companies are seeking to avoid being overtaken by emerging Internet services, and so it makes good corporate sense for them to become involved in ownership of online services.

As major media corporations become involved in online content delivery, those Internet content providers lacking the financial and content resources to compete find themselves marginalized on the Web. Furthermore, as traditional media giants extend their reach into the Internet, they are able to use Internet sites as promotional support both for other online material and for their broadcast, film, or video activities.

The end result is a corporatized internet promoting the goods, services, and values of major companies and their subsidiaries. Links on such pages are to other sites that typically are within the corporate family, and the material and advertising create a closed world of cross-promotional corporate content. The corporate term for this practice is “synergy.” However desirable synergy may appear to the corporate planners, from social and aesthetic points of view, such circular linkages limit content and choice.

New Organizations

News sites routinely incorporate multimedia materials into their Web sites. Traditional text and images are often liked to video or audio links providing additional commentary on news stories. Such multimedia linkages may bring together previously diverse corporate divisions, as in the case of the BBC news Web page (http://news.bbc.co.uk). Linkages to audio formats from BBC radio and video from the television division combine with text to provide different dimensions of the story derived often from differing styles of reporting.

The video used on such sites is generally in a streaming video format and thus has low frame rates compared with television. The videos are often jerky and with poor resolution. Nonetheless, the use of this streaming video technology is becoming more common in reporting from the field and was often used during the 2002 U.S. invasion of Afghanistan. Using cellular phones or satellite uplinks, reporters with personal computers and Web cams presented live video reportage from places with insufficient infrastructure and access to earth stations to support standard satellite uplinks.

Such reporting practices make clear that, just as the personal computer and the Internet made publishing a possibility for anyone with the equipment, so other activities, such as video reporting, are now technologically available to large segments of the population. The video camcorder spawned both new dimensions to news reporting, with amateur footage being used in news broadcasts of local, national, and international significance. From house fires to the Rodney King beating to the impact of the second plane into the World Trade Center on September 11, 2001, amateur video is becoming a significant supplier of news. Such content conveys the illusion that local or individual citizens have input into media narratives or forms, however. The opposite is increasingly true: Mass media create the aesthetics for home video and for much of multimedia.

The Entertainment Industries

In addition to news organizations, multimedia has allowed the development of numerous new corporate forms of the entertainment industry. One of the most significant corporate groups in pioneering new multimedia technologies and in commercializing the online services is the pornography and erotic materials industry. These companies were among the first to deploy secure payment systems and to use live Web cams and real-time streaming video. Adult sites draw significant percentages of Internet traffic and generate large amounts of profit, although estimates of the amount they contribute to total traffic are contested and difficult to verify. Nonetheless, it is clear that much of the popular technology of multimedia has been driven by the adult entertainment industry.

The future of the pay-per-view industry is probably foreshadowed in the history of the pornographic film deliveries. Recently, sales and rentals of adult erotic videos began to plateau, and on-demand television rentals and online rentals continued to grow. As quality and speed of delivery improve, it is likely that many people will begin to order other forms of video in this way rather than renting or purchasing VCR or DVD copies. With the penetration of DVD home recording, such delivery systems will grow in popularity as the transference of video from cable to homemade DVDs becomes as common as the burning of CDs. As with VCRs, MP3, and CDs, the use of home DVDs well undoubtedly trigger a round of debate and litigation regarding copyright issues.

Telecommuting and Teleconferencing

Telecommuting and teleconferencing have suffered from the same problems as other business areas: poor bandwidth, complex software, and a lack of consumer familiarity with the technology. Just as these barriers are being hurdled in video delivery, increased simplicity in software and growing high-speed connectivity are likely to make these uses of digital technology more commonplace. As video and audio quality in real time approach television standards, teleconferencing with video will become increasingly popular and more widely used, especially in business and education. Furthermore, many corporations are beginning to consider the need to incorporate broadband infrastructure into building design and renovation as the horizon of these technologies moves closer.

Many of the applications of converged video and text are already in growing use in education. Traditional methods of distance or mixed delivery mode course provision are easily supplemented with multimedia materials, either online or delivered on CD. Not only have remote classes, taught over CCTV, become more common, but with improvements in streaming video quality, two-way interactive classes taught over the Internet are likely to also grow in popularity. The screen compositions in such online classes increasingly feature integrated sound, text, and video. The trend toward this form of distance education is being fueled by a shrinking pool of traditional students, and growing competition among domestic and overseas universities. The technology is proving integral to the widespread Anglo-American establishment of degree programs overseas, particularly in Asian countries.

SOCIAL EFFECTS

Convergences of technology allow people to conduct conversations, meetings, classes, and other social activities even when they are not physically in the same place. They also challenge traditional ideas of what constitutes a text, how knowledge functions, and raise issues regarding the differences between the represented world of text and media and the physical world.

Glocalization

Glocalization is a contraction of “global” and “local” and indicates that a virtual environment may have some of the characteristics of being both decentralized and still connected to some central hierarchy. For example, telecommuting and other changes in workplaces undermine the traditional separations between home and office. Such changes have consequences for both the worker and for the company. The culture of the office or corporation, for example, is difficult to establish and maintain if people do not share the same physical environment and have some face-to-face interaction with one another.

At the level of the social, glocalization may contribute to the de-realization of the world. As representations of the world grow in number and levels of verisimilitude, the tendency is to accept that the representations are the world. In other words, the growth of virtual worlds may contribute to what Jean Baudrillard called “simulation,” that is, a social system of images and words that is endlessly self-referential without having any reference to a “real” world. In a similar vein, Paul Virilio (1997) writes that the virtual world displaces the sense of the real world, substituting itself for the real. The political and social implications of these two positions are numerous.

Increasing simulation of the real or providing a substitute for it may encourage people to alter the ways that they interact with the world and with other people. Online gaming, cybersex, and chat rooms have already shown that traditional social markers such as gender, race, and age are frequently changed by people who meet only in a digital environment. Identity itself may become more fluid as people change their digital persona to fit situations without reference to a known or widely recognized core self. As a consequence, virtues such as honesty and integrity, become more contextual and transitive. The digital world does not link identity the uniqueness of given localities in the physical world.

“Data Smog”

Data smog refers to the problem of sorting meaningful information from increasingly large amounts of material with which one is presented (Schenk, 1997). Visual and audio effects may add little of use to messages in some cases and may tend simply to create large amounts of noise or “data smog” by inflating information amounts without adding to meaning, context, or content. The proliferation of information both on a social level and at the level of the individual message create problems for the reader. It may be difficult to sort through the volume of information present to determine what is of interest and importance.

Too much information on a single page overwhelms the viewer or reader. It also increases download times and produces congestion, especially on intranets and wide area networks (WANs). The end result may be that the information page is too cluttered to be easily accessed or navigated. Without simple guides and menus, a page may require too much time to evaluate. This is especially true with multimedia pages that often reflect a greater interest in displaying the "bells and whistles" of the new technology rather than serving the needs and interests of the viewer.

At the social level, it is difficult to sort important information from the background of available materials. Traditionally, the organization of news and information was the job of editors, publishers, and librarians. They performed the gate-keeping function of selecting and presenting information. With the number of publications available online, however, additional filters and strategies for organizing materials are often needed. One consequence of the amount of available materials online is the growing sophistication and use of search engines with which people may locate their own news reports on topics of interest to them rather than relying on traditional news organizations. To date, most people continue to use the same news sources online as they do otherwise. Nonetheless, niche publications have proliferated as many costs associated with printing and distribution do not arise in online production.

Continuation and Extension of the Digital Divide

The distinctions between the information haves and the have-nots continue to expand and multiply. Larger numbers of people find themselves excluded from each new technology by their inability to afford them. New technologies tend to be expensive, and even after economies of scale have made technologies more inexpensive, they still remain beyond the reach of many. Poorer persons and even whole countries may lack the resources to afford the hardware and the infrastructure required by new technological developments, and as such may be excluded from participating in developing the forms and uses of new technologies.

Furthermore, as some companies and countries automate services such as banking, education, and access to government services, people without the technology are excluded. Local banks may close branches, government forms and documents may be unavailable, products may be unobtainable, and educational opportunities may be lost if such radical monopolies of digital culture replace traditional ways of providing goods and services within the community. Local services, such as loan offices, schools, and government departments, are easily transferred to remote offices that have no local knowledge of individual people's lives and circumstances. In such ways convergent technologies can contribute to the corporatization of everyday life by eliminating face-to-face interaction and the possibility of people's doing things together.

EDITING AND CREATING NEW TEXTS

New Textualities

Some writers have suggested that the combination of data, video, text, and other elements in a single digital form creates new forms of textuality and so changes the ways in which readers approach and use texts. Another way of stating this is to say that the new technologies allow textual forms to emerge which have been described and theorized in critical theory and some branches of linguistics, particularly poststructuralist theory. The long-term implications of these new textualities are not yet clear. On one hand, such new, multimedia textual forms may open up techniques for expression that are not possible with standard printed texts. On the other hand, new textual forms may contribute to the demise of certain ways of reading and writing, and with them, the focused, concentrated attention and logic of thinking for which they acted as templates may also be lost.

Printed texts are generally read in a linear way, and the traditions of reading developed around logical patterns of exposition, narrative, and explanation. Digital and multimedia texts, particularly when combined with hypertext forms allow and even encourage the reader to follow nonlinear lines of argument that extend beyond the text at hand and that may encompass other texts, images, sounds, and so on. As such, the convergence of diverse kinds of representation may change traditional ways of reading as well as traditional attitudes toward the text. Ideas such as authorship and authorial control of a work require redefinition under these new conditions in which there may be no single author or group of authors and the text itself may incorporate numerous texts. In similar ways, the current definitions of intellectual property and copyright are already proving inadequate for some emerging forms of text.

Certainly the designing, editing, and reading of multimedia require new sets of skills and knowledges, and these change both how the text is read and what kinds of knowledge are extracted from the multimedia text. Images and video, although seeming to present more complete information than text, are problematic. The selection of particular images, issues of video editing, and the contexts in which these images appear all influence the meanings that are taken from them. Yet images have an immediacy and power that text lacks. Additional strategies for critical reading of multimedia are required that would allow readers to evaluate the function and operation of images and video and to make informed decisions regarding the validity and use of those images.

Most people, however, tend to read multimedia pages using techniques learned from film, television, and reading, and these are largely the products of mass-media corporations and are designed to appeal to large audiences. As such, the liberating potential that some see for the new textualities remains unrealized. The skills that one learns from mass corporate media are largely the skills required to consume more media as well as ancillary products which are advertised in cross-marketing schemes. The limitation of media literacy to mass-media and corporate forms constitutes a kind of gate keeping that establishes the norms of representation. Whatever does not fit both in narrative and in format declines in status and popularity. Other gate-keeping functions of print-based text, such as a known publisher, a physical form for the text, and literate conventions regarding how one reads, continue to help determine the ways in which texts are understood, fostering the expectation that texts contained a single or preferred meaning. This stability constituted an implied contract between the author and the reader that certain conventions would be followed that would lead to certain kinds of understandings of the text.

Nonetheless, as the skills and technologies for generating multimedia become more widely available and used, a democratization of skills will follow, just as printing produced the amount of materials necessary for general literacy. As people learned to read, they also began to write and to use that technology in unanticipated ways, creating new kinds of writing. Just as writing became a general skill, differentiated from reading, in early modern times, audio and video editing, combined with HTML composition, may well become standard skills. Already, Web sites exploring alternative approaches to organizing and arranging multimedia are demonstrating possibilities for multimedia outside of the conventions of advertising, film, and traditional aesthetics.

GLOSSARY

Digital subscriber line (DSL) A technique for transmitting broadband digital data over paired copper telephone lines.

Motion picture expert group (MPEG) A group which sets standards for compressing digital video data to reduce the size of the file.

Multi-protocol label switching (MPLS) A protocol for attaching a label to a data packet that contains information regarding its routing and priority and other information.

Phatic communication An exchange between human beings that conveys little apparent content and the primary purpose of which is to establish or to acknowledge mutual recognition.

SOHO An acronym for small office, home office.

Quality of Service (QoS) A determination that a particular transmission is satisfactory for the users.

Voice over Internet protocol (VOIP) The digitizing of voice, such as telephony, and its transmission over Internet servers to a receiver.

CROSS REFERENCES

See *Interactive Multimedia on the Web; Local Area Networks; Multimedia; Public Networks; Telecommuting and Telework; Voice over Internet Protocol (IP); Web Quality of Service; Wide Area and Metropolitan Area Networks.*

REFERENCES

Baudrillard, J. (1983). *Simulations*. New York: Semiotext(e).

The Benton Foundation (1996). *The Telecommunications Act of 1996 and the changing communications landscape*. Retrieved November 19, 2002, from http://www.benton.org/Library/Landscape/landscape.html

Carey, J. W. (1991). *Communication as culture* (2nd ed.). New York: Routledge. (Original work published 1989.)

Ellul, J. (1964). *The technological society* (J. Wilkinson, Trans.). New York: Vintage Books.

ITU (2003). Free Statistics Homepage. Retrieved April 2, 2003, from http://www.itu.int/ITU-D/ict/statistics/at_glance/basic02.pdf

Luhman, N. (2000). *The reality of the mass media* (K. Cross, Trans.). Stanford, CA: Stanford University Press.

Schenk, D. (1997). *Data smog*. London: Abacus.

Virilio, P. (1997). *Open sky* (Julie Rose, Trans.). London: Verso.

FURTHER READING

The Artmuseum (n.d.). *Multimedia*. Retrieved on Spetember 20, 2002, from http://www.artmuseum.net/w2vr/project.html

Barrett, E., & Redmond, M. (1995). *Contextual media: Multimedia and interpretation*. Cambridge, MA: MIT Press.

Barthes, R. (1970). *Writing degree zero* (A. Lavers & C. Smith, Trans.). Boston: Beacon Press.

Brittain, P., & Farrel, A. (2001). *MPLS traffic engineering: A choice of signaling protocols*. Retrieved November 22, 2002, from http://www.dataconnection.com/downloqad/crldprsvp.pdf

Butzgy, M. (2002). *Writing for the multimedia: A guide*. Retrieved November 20, 2003, from http://home.earthlink.net/~atomic-rom/contents.htm

Egan, B. L. (1996). *Information superhighways revisited: The economics of multimedia*. Boston: Artech House.

Government Accounting Office (2001). *Telecommunications: Characteristics and choices of Internet users* (GAO 01–345). Washington, DC: Author.

International Engineering Consortium (2002). *A comparison of multiprotocol label switching (MPLS) traffic-engineering initiatives*. Retrieved January 15, 2003, from http://www.iec.org/online/tutorials/mpls-traffic/index.html

International Telecommunications Union (2001). *ITU telecommunication indicators update 2001*. Retrieved May 8, 2003, from http://www.itu.int/journal/200102/E/html/indicat.htm

International Telecommunications Union (2002). *ITU Internet reports 2002: Internet for a mobile generation*. Retrieved May 8, 2003, from http://www.itu.int/osg/spu/publications/mobileinternet/chapter1.html

McChesney, R. (1999). *Rich media, poor democracy: Communication politics in dubious times*. Urbana, IL: University of Illinois Press.

Rheingold, H. (1993). *Virtual communities*. New York: Addison-Wesley.

The UCLA Internet report (2001). Surveying the digital future: Year two. The UCLA Center for Communication Policy. Retrieved September 20, 2002, from http://www.ccp.ucla.edu/pdf/UCLA-Internet-Report-2001.pdf

Virilio, P. (1991). *The aesthetics of disappearance*. New York: Semiotext(e).

Copyright Law

Gerald R. Ferrera, *Bentley College*

INTRODUCTION

E-commerce transactions create unique copyright issues as digital content is transmitted over the Internet. Digital content may take the form of words, videos, music, and terms of use and privacy policies that are commonly found on Web sites. Provided this digital content is an "original work," it is the appropriate subject matter of copyright protection and federal registration in the United States Copyright Office.

Because federal copyright law provides the owner with exclusive rights, primarily to reproduce the original work, copyright ownership incurs a special significance on the Internet where Web site design and development are costly ventures and valuable assets to the e-business. Because online content is digitized, the user has the ability to reproduce and send it immediately to countless others in violation of the copyright owner's exclusive right to reproduce the work. Software programs, such as Lotus Notes, allow a user to send e-mail with attached copyrighted documents that are reproduced in perfect form and may be sent on innumerable occasions to a vast global audience. This poses special and unique problems for copyright protection on the Internet.

Federal copyright law is found in the Copyright Act [17 U.S.C. 101–1205], which protects the authors (or transferees and licensees of copyright material) from copyright infringement for "original works of authorship fixed in a tangible medium of expression" [17 U.S.C. Sec. 40l (d)]. A licensee of a copyrighted work infringes a copyright by exceeding the terms and conditions of the copyright license (see *S&H Computer Sys, Inc. v. SAS Inst., Inc.*, 1983). Although copyright registration is not required in the United States, there are advantages to federal registration in the U.S. Copyright Office. Registration does not confer ownership of a creative work, but rather statutory benefits. Among these benefits are (a) access to the federal court system, unless the copyright has been registered a court may dismiss an action for infringement [17 U.S.C. sec. 411 (a)], (b) statutory damages in lieu of actual damages that may be difficult to prove [17 U.S.C. sec. 412 (2)], and (c) presumption of the validity of the copyright [17 U.S.C. sec. 410 (c)]. Depending on the nature of the e-business, federal registration may be more suitable for some companies than for others. Because the original content on Web pages qualifies for federal registration of the site in the U.S. Copyright Office, one of the first orders of business for an e-company should be to consult with legal counsel concerning the advantages of filing its Web pages with the U.S. Copyright Office for federal copyright registration.

COPYRIGHT AS INTELLECTUAL PROPERTY

Copyright is only one piece of intellectual property historically comprising four pillars: copyrights, trademarks, patents, and trade secrets. The first English copyright act, the 1710 Statute of Anne was the origin of U.S. copyright law. The U.S. Constitution provides the foundation for copyrights found in Article I, Sec. 8, clause 8, that states in part "Congress shall have the Power to promote the Progress of Science anduseful Arts, by securing for limited

Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries." Hence, the authors of original works of authorship have the right to preclude others from the use of those works during the limited time determined by Congress. In a significant case, *Computer Assoc. Int'l. v. Altai, Inc*., 982 F.2d 693, 696 (2d Cir. 1992), the court stated that

> the copyright law seeks to establish a delicate equilibrium. On the one hand, it affords protection to authors as an incentive to create, and, on the other, it must appropriately limit the extent of that protection so as to avoid the effects of monopolistic stagnation. In applying the federal act to new types of cases, courts must always keep this symmetry in mind.

Copyright law provides a balance between the public's use of the work and the granting of exclusive rights to the author. In *Suntrust Bank v. Houghton Mifflin Co.* (2001) the court stated, "the Copyright Act promotes public access to knowledge because it provides an economic incentive for authors to publish books and disseminate ideas to the public." The exclusive rights granted to copyright owners is limited for a fixed duration, however.

Duration of a Copyright

In October 1998, Congress enacted the Sonny Bono Copyright Term Extension Act. The Act extends the time of most copyrights for 20 years. Copyright material created by individuals was previously for the life of the author plus 50 years and is now for the life of the author plus 70 years. After that time, the copyright falls into the category of public domain for the benefit of the common good where the work may be reproduced without copyright infringement. Corporations owning copyrights were also granted an additional 20 years from the 75-year period for works created for hire, bringing the term for corporations to 95 years from the date of the first copyrighted work. The extension of the term for 20 years is consistent with the copyright law granted by the European Union. Copyright durations have been amended over the years and the applicable duration will vary depending on the facts of each case.

Copyright Duration Challenged: *Eldred v. Ashcroft*

The United States Supreme Court has decided the case of *Eldred v. Ashcroft* (2003) on the issue of whether the constitutional limits of the congressional power to extend the term of copyright has been exceeded as provided in the Copyright Clause. Congress has the authority to issue copyrights "for limited times" to "promote the progress of science and useful arts." This copyright protection is a monopoly giving exclusive rights to the copyright owner with some limitations for fair use during the copyright period. The basis for that entitlement is to give the authors incentive to write and publish material and at the same time to promote the progress of the arts. One could argue as our cultural heritage is transmitted from one generation to the next, it is necessary at some point for this copyright protection to end and for the copyrighted work to fall into the realm of public domain. The retroactive extension for 20 years granted in the Sonny Bono Copyright Term Extension Act of 1998 has a delaying effect on the public domain's reception of expired copyrighted works. The right of Congress to grant a monopoly to copyright owners is subject to the limited time provision of the U.S. Constitution to ensure a vast public domain of free works available for the common good. However, in upholding the congressional power to extend the copyright term in the *Eldred* case the U.S. Supreme court ruled

> Copyright purpose is to promote the publication of free expression. The Framers intended copyright itself to be the engine of free expression. By establishing a marketable right to the use of one's expression, copyright supplies the economic incentive to create and disseminate ideas. As we read the Framers' instructions, the Copyright Clause empowers Congress to determine the intellectual property regimes that, overall, in that body's judgment will serve the ends of the Clause. The wisdom of Congress' action, however, is not within our province to second guess.

Works Made for Hire

A work for hire is a work specially created by an employee within the scope of his or her employment that is specially ordered or commissioned under the Copyright Act, sec. 101, discussed later in this chapter. For instance, a creative employee working for an e-business may produce, develop, or upgrade a Web site for the company. Can the employee later claim a copyright interest in the creative work? The answer to that question depends on whether the employee was acting within the scope of his or her employment or the work was expressly or implicitly found to be a work for hire or commissioned works; if so, the company would own the copyrighted work and would be considered the author for copyright purposes. In most instances, the employer owns the creative works of its employees. The status of the individual creating the copyrighted work is determined under the agency principles relative to an employer–employee relationship. The courts will look at the degree of supervision and control exercised by the employer over the employee in creating the work.

A checklist based on a test determined by a famous Supreme Court case, *Community for Creative Nonviolence v. Reid* (1989), commonly referred to as the "Reid" test, provides guidelines for ascertaining who owns the copyright in the creative work. The court considered:

- The employer's right to control the manner and means of the employee's work (the more control exercised by the employer, the more likely the creative work is a work for hire and owned by the employer)
- The level of skill required to create the work
- The sources of tools and instruments necessary to create the work

- The location of the work
- The duration of the relationship between the employer and the employee
- The employer's right to assign additional projects to the employee
- The method of payment
- The tax treatment of the relationship of the parties
- Whether the project is part of the regular business of the employer

A company should consider a discussion with its attorney about amending the employee's job description so that it includes the production by the employee of original and creative works that may be copyrighted and further that those works are owned by the employer. The employment agreement should provide for the assignment of the employee's works to the employer in the event that the creative works are not works for hire.

Selectively Ordered or Commissioned Works

Section 101 of the Copyright Act defines a work for hire as

> a work prepared by an employee within the scope of his or her employment or a work specially ordered or commissioned for use as a contribution to a collective work, as a part of a motion picture or other work, as a translation, as a supplementary work, as a compilation, as an instructional text, as a test, as answer material for a test, or as an atlas, if the parties expressly agree in a written instrument signed by them that the work shall be considered a work for hire.
>
> If a company hires an independent contractor, it must have a written contract expressly stating that the specially ordered or commissioned works should be considered a work for hire.

Registration in the U.S. Copyright Office

The moment the original work is created and "fixed in a tangible medium of expression," copyright law protects it, even if it is never federally registered (see *Data Gen. v. Grumman Sys. Support Corp.*, 1994: "copyright protection attaches the day original expression is fixed in a tangible medium"). Although copyright law protects an original and creative work, filing for the registration in the U.S. Copyright Office is a prerequisite for initiating a legal claim for copyright infringement in the federal court system. The courts have held that "[t]he registration requirement is a jurisdictional prerequisite to an infringement suit" (*M.C.B. Homes, Inc. v. Ameron Homes, Inc.*, 1990).

Plaintiff's Burden of Proof in a Copyright Infringement Case

A person or company bringing a copyright suit must prove that it is the owner of the copyright and that an unauthorized copying has occurred in violation of an exclusive right and a specific theory of copyright infringement against the defendant. In *Computer Assocs. Int'l v. Altai, Inc.* (1992), the court ruled that "[i]n any suit for copyright infringement, the plaintiff must establish its ownership of a valid copyright, and that the defendant copied the copyrighted work." A defendant may be primarily liable as a direct infringer or secondarily liable as a person or company that contributes to the infringement in some manner or a vicarious infringer. These theories of copyright infringement will be discussed later.

Subject Matter of Copyright

The subject matter of what is appropriate material for copyright protection is vast. It includes "literary, musical, dramatic, and choreographic works, pantomimes, and pictorial, graphic, or sculptural works, including the individual images of a motion picture or other audio visual works" (U.S. Copyright Act, §1027). Web page content often includes literary, pictorial, graphic works, and images of motion pictures or audiovisual works that are all the appropriate subject matter of copyright protection.

Exclusive Rights of the Copyright Owner

When federal registration is obtained, the copyright owner is granted exclusive statutory rights found in section 106 of the Copyright Act as follows:

- The right to reproduce the work
- The right to prepare derivative works based on the original material (in *Synercom Technology, Inc. v. University Computing Co*. 1978, the court ruled as follows: "[I]t is as clear an infringement to translate a computer program from, for example, FORTRAN to ALGOL, as it is to translate a novel or play from English to French. In each case the substance of the expression (if one may speak in such contradictory language) is the same between original and copy, with only the external manifestation of the expression changing. Likewise, it would probably be a violation to take a detailed description of a particular problem solution, such as a flow chart of step-by-step set of prose instructions, written in human language, and program such a description in computer language.")
- The right to distribute copies of the work to the public by sale, rental, lease or lending
- The right to perform the work publicly, (in the case of literary, musical, dramatic, and other such types of work)
- The right to display the copyrighted work publicly (in the case of literary, musical, graphic, sculptural and other such types of works)
- The right with sound recordings to perform the work publicly by means of digital audio transmission

Web pages can be given the same copyright protection as books, music, and movies, preventing others from reproducing the work without express permission from the copyright owner. There are limitations on the copyright owner's exclusive rights that will be addressed later.

Linking

Web users are accustomed to link from one site to another in their search for reliable information. Web site operators commonly link pages without permission of the owner on the assumption that posting a site on the Internet grants an implied license to link it to another Web site. They would argue that, because there is no copying of the content but rather a connection to the site, there is no copyright violation.

In a federal court case, *Intellectual Reserve, Inc. v. Utah Lighthouse Ministries, Inc.* (1999), copyrighted documents from the Church of Jesus Christ of Latter-Day Saints containing infringed materials were knowingly posted on defendant's Web site without authorization from the owner. The court ruled that

> when a person browses a Web site, and by so doing displays the Handbook, a copy of the Handbook is made in the computer's random access memory (RAM), to permit viewing of the material, and in making a copy, even a temporary one, the person who browsed infringes the copyright.

Based on the theory that posting of the links on defendant's Web site to the copyrighted documents facilitated infringements by Web users, the court found the owner of the linking site liable for contributory infringement (discussed below) and granted a preliminary injunction ordering the defendant to remove the infringing materials. Web site managers should consult with their lawyers on links to sites that may contain infringing material and perhaps consider obtaining permission to link from the copyright owner.

Inline Linking

In *Kelly v. Arriba Soft Corp.* (2002), the plaintiff, Leslie Kelly, owner of copyrighted photographs, sued the defendant, the operator of a search engine that displayed "thumbnail" or small pictures of Kelly's works, for copyright infringement. The defendant, without permission, lifted the photographs directly from Kelly's Web site and then displayed them on its search engine. This practice infringed upon Kelly's exclusive right to "display the copyright work publicly." The court ruled as follows:

> Arriba is directly liable for infringement. Arriba actively participated in displaying Kelly's images by trolling the web, finding Kelly's images, and then having its program inline link and frame those images within its own web site. Arriba acted as more than a passive conduit of the images by establishing a direct link to the copyrighted images. Therefore, Arrriba is liable for publicly displaying Kelly's copyrighted images without his permission.

This case is a reminder that search engine operators should have their legal counsel review how it obtains its database of materials to assure copyright compliance.

Originality of the Work

The statutory requirement that the author produce an "original work" is not in the same category as the originality of a patent. Mere factual material, however, is not copyrightable unless it is accompanied by explanatory text. This raises the issue as to whether a database is a copyright production.

Databases

Because it is a fundamental legal principle that only "original" works may be copyrighted, the question arises whether the material found in a database qualifies for copyright protection. When Web sites use "cookies" to collect click-stream information from users and compile that information in their database for possible sales to marketers, the question arises as to whether the database is copyrightable material.

Providing the material in the database is not merely factual representations but rather the author's creative selection and arrangement of the data, it may be copyrightable. In a famous copyright infringement case, *Feist Publications, Inc., v. Rural Telephone Services Inc*. (1991), the defendant was the publisher of a telephone directory that reproduced more than 1,000 of the plaintiff's telephone numbers without its consent. The court held that the plaintiff's mere arrangement of factual material lacked originality because

> facts, whether alone or as part of a compilation, are not original and therefore may not be copyrighted. A factual compilation is eligible for copyright if it features an original selection and arrangement of facts, but the copyright is limited to the particular selection or arrangement. In no event may copyright extend to the facts themselves.

If factual information about a consumer is compiled in a manner that the author selects as optional features in a unique fashion, the material may be copyrighted. In a federal court case, *CCC Information Services, Inc., v. McAllen Hunter Market Reports, Inc*. (1994), the facts involved the "Red Book" that listed the retail value of used automobiles. The court held that the listing in the defendant's Red Book were original because the Book made an adjustment for mileage in 5,000-mile increments and used the concept of an "average" vehicle as the subject matter of its evaluation.

Other issues regarding databases used and reproduced by online service providers pertain to information aggregators on the Internet. Can an online service provider of information legally protect that information from redistribution by another online service provider? A recent case between eBay, Inc. and Bidder's Edge, *eBay, Inc. v. Bidder's Edge, Inc*. (2000), involved a program whereby Bidder's Edge allowed consumers to request specific products at a specific price by searching Web sites. Bidder's Edge "crawlers" accomplish this copying of product information from eBay and other Web sites. Consumer could access this product information on the Bidder's Edge Web site. eBay claimed use of the content collected by Bidder's

Edge constituted an unlawful trespass on the eBay computer system. The court issued a preliminary injunction against Bidder's Edge restricting it from accessing the eBay site.

Because information aggregators provide important information to consumers, the courts will continue to seek a balance between the content ownership of data (e.g., in a database) and the reproduction of that information. Database access and use restrictions will continue to be addressed by the courts and the legislature.

THEORIES OF COPYRIGHT INFRINGEMENT

Exclusive statutory rights granted to a copyright owner are protected in the federal court system. Once a copyright owner establishes in court that it is the owner of the original work and that the defendant has infringed on that work by copying it in some fashion without the owner's permission, then the plaintiff/owner must proceed by proving the case under an established theory of copyright infringement. The theories of copyright infringement are not stated in the Copyright Act but have been developed by the courts and are now part of our copyright law. There are three theories of copyright infringement: direct, contributory, and vicarious.

Direct Infringement (Primary Liability)

A direct infringer is the person, company, or agent that violates one of the copyright owner's exclusive statutory rights. An online newsletter using copyrighted information without the author's permission and posting it online would be a direct infringer. Direct infringement is also known as primary liability.

Contributory Infringement (Secondary Liability)

Contributory infringement is similar to the liability assessed to a person who assists another in committing a tort in the civil law. One does not have to be the direct actor in a wrongful act to be legally responsible and accountable in a court of law for the loss to the defendant. There is, however, a prerequisite that a contributory infringer must have *knowledge or have reason to know of the direct infringement* to establish liability. Once that fact is established, a party "with knowledge of the [direct] infringing activity, induces, causes, or materially contributes to the infringing conduct of another, may be held liable as a contributory infringer" (*Gershwin Publ'g Corp. v Columbia Artist Mgmt., Inc.*, 1971).

In a famous U.S. Supreme Court case, *Sony Corporation of America v. Universal City Studios, Inc.* (1984), Sony, which manufactures and sells home videotape recorders, was sued by Universal, which owned copyrights on some of the television programs broadcast on public airwaves. The public used Sony's videotape recorders to record some of the broadcast. The court stated, "This practice, known as 'time-shifting,' enlarges the television viewing audience." In refusing to find contributory liability on behalf of Sony, the court noted that

> contributory infringement has been applied in a number of . . . copyright cases involving an ongoing relationship between the direct infringer and the contributory infringer at the time the infringing conduct occurred. . . . The only contact between Sony and the users of the Betamax . . . occurred at the moment of sale. . . . One may search the Copyright Act in vain for any sign that the elective representatives of the millions of people who watch television every day have made it unlawful to copy a program for later viewing at home, or have enacted a flat prohibition against the sale of machines that make such copying possible.

Knowledge of the infringement may be based on the direct perception of the contributory infringer or on the knowledge that *a reasonable person in a similar set of circumstances* would have. In the now famous *A&M Records, Inc., v. Napster, Inc.* (2001), the U.S. Court of Appeals found Napster to be secondarily liable for copyright infringement based on the system it provided that allowed the users to download copyrighted music. The court ruled that Napster had actual knowledge that specific infringing material was available on its system and that it could have blocked access to this system and failed to do so, and this conduct materially contributed to the direct infringement.

The standard is whether a reasonable person, in that environment, would have been aware of the direct infringement. The doctrine of contributory infringement may pose special problems to an online business venture because the traditional way of limiting personal liability would be to incorporate the electronic business. If the directors, officers, stockholders, or venture capitalists that funded the business knew or had reason to know of the direct infringement and allowed it to continue, it is arguable that they could be personally sued for contributory infringement and not have the benefit of limited liability provided by corporate law.

For example, suppose the investors, stockholders, and corporate directors all knew that the program and business model used on the company Web site allowed others to download copyrighted material. It is possible under the doctrine of contributory copyright infringement they could all be jointly liable to the copyright owner.

Vicarious Infringement

Vicarious infringement, similar to contributory infringement, has its origin in the common law based on court decisions. The doctrine of vicarious copyright liability is established on the theory that the defendant has the right and ability to supervise the direct infringement and also has a direct financial interest in the infringed copyrighted work. This test does not require active supervision. Vicarious liability is found in the ancient employment doctrine of *respondeat superior* based on the theory that an employer is responsible for the carelessness of its employees while they are carrying on business activities. Provided that the employees are acting within the scope of their employment contract, the employer has the right to supervise and control employees' activities and is legally responsible for their wrongful acts causing injuries to others. Vicarious liability also includes independent contractors who have the right and ability to control others and maintain a direct financial interest in the copyrighted material.

An e-business could be liable for copyright infringement if one of its employees violated a copyright law while engaging in company business. Vicarious copyright infringement liability does not require an employer–employee relationship. Liability is based on the fact that the vicarious infringer had the right to control and stop the direct infringing and, in addition, that he or she received a financial benefit from the infringement.

DEFENSES TO COPYRIGHT INFRINGEMENT

Copyright Law was intended to establish a balance between the exclusive rights of authors to use for a limited period of time their original works and at the same time, under limited circumstances, to provide the public with that information. This is a recognition by Congress that our copyright policy, although it grants an exclusive use to the copyright owner, also allows that information, within limited circumstances, to flow for the benefit of the public good. One of the most important doctrines found in the Copyright Act that allows for this process is the theory of fair use.

Fair Use

Certain infringements are excused from liability when the use is "fair." Fair use had its origin in case law long before it was adopted by Congress and became part of the U.S. Copyright Act. Judges recognized that in limited circumstances the general public should have the right to reproduce copyrighted material for the benefit of the common good. As early as 1995 in *Bateman v. Mnemonic, Inc*., the court noted,

> Originally, as a judicial doctrine without any statutory bases, fair use was an infringement that was excused—this is presumably why it was treated as a defense. As a statutory doctrine, however, fair use is not an infringement. Thus, since the passage of the 1976 [Copyright] Act, fair use should no longer be considered an infringement to be excused, instead, it is logical to view fair use as a right.

In *Steward v. Abend* (1990), the U.S. Supreme Court ruled the purpose of the fair use doctrine is to "avoid rigid application of the copyright statute when, on occasion, it would stifle the very creativity that the law is designed to foster." That judicial theory has been implemented in the Copyright Act in section 107 as a limitation on the exclusive rights of the copyright owner. The burden of proof is on the defendant who claims the affirmative defense of fair use and must establish that the copying was used within the statutory parameters set out in section 107 that states, "the fair use of a copyright work, including such use by reproduction in copies or phonocopies or by any other means . . . for purposes such as criticism, comment, news reporting, teaching . . . scholarship, or research is not an infringement of copyright." The statute enumerates four nonexclusive factors that make up the doctrine of fair use and the courts must consider all of them in each case:

> (1) the purpose and character of the use including whether such use is of a commercial nature or is for non-profit educational purposes; (2) the nature of the copyright work; (3) the amount and substantiality of the portion used in relation to the copyright work as a whole and (4) the effect of the use upon the potential market for or value of the copyright work.

Nonprofit educational purposes allow these institutions to use copyrighted works within a limited framework. State corporate law provides authorization for an enterprise to be organized as a nonprofit corporation. Federal and state eligibility for a nonprofit results in a two-way tax proposition useful for these ventures. The person contributing to the nonprofit is allowed to take a tax deduction, and the nonprofit receives that income tax free. There are countless numbers of Web sites that are nonprofit entities. If those Web sites were to display copyright priority over copyright material and the other factors are satisfied, the defendant may be able to argue fair use over the objection that the publication constituted a copyright infringement. The fact that a company is a nonprofit and is using copyrighted material is not conclusive evidence that it fits within the doctrine of fair use. Social expression and the free flow of ideas, especially within an educational context, takes priority over a copyrighted work. The court ruled in *Universal City, Inc. v. Reinerdes* (2000) that the [fair use] "doctrine . . . limits the exclusive rights of a copyright holder by permitting others to make limited use of portions of the copyrighted work, for appropriate purposes, free of liability for copyright infringement." For example, it is permissible for someone other than the copyright owner to reprint or quote a suitable part of a copyrighted book or article in certain circumstances. The doctrine traditionally has facilitated literary and artistic criticism, teaching and scholarship, and other socially useful forms of expression. The courts have viewed it as a safety valve that accommodates the exclusive rights conferred by copyright within the freedom of expression guaranteed by the First Amendment (*New Era Publ'ns Int'l, ApS v. Henry Holt & Co., Inc*., 1989: "the fair use doctrine encompasses all claims of first amendment in the copyright field").

First Sale Doctrine

The First Sale Doctrine, found in section 109 of the Copyright Act, limits the copyright owner's exclusive right to distribute publicly a copy of the work providing the copyright material was lawfully acquired by another person. Under the Copyright Act, "the owner of a copy is entitled, without the authority of a copyright owner, to sell that copy" (sec. 19). The purchaser of a copyrighted work can generally sell that material to another without violating the copyright law. One must be cautioned, however, that if there is a licensing agreement of a copyrighted work, the copyright owner under that agreement could retain the right of distribution. For example, software may be downloaded under a licensing agreement whereby the computer software company may continue to own the copyrighted information on the program while the licensee maintains use of the program consistent with the licensing agreement. The courts are split on the issue of whether a user manifests assent to a licensing agreement merely by opening a software package and using the software (cf. *Step-Saver Data Sys., Inc. v. Wyse Tech., Inc.*, 1991 [no shrink-wrap license found by opening the box] and *ProCD, Inc. v. Zeidenberg*, 1996 [shrink-wrap license found binding on the user]).

REMEDIES FOR COPYRIGHT INFRINGEMENT

The copyright owner can seek various remedies under the Copyright Act. These include injunctive relief in which the court may force the infringer to stop the copying and may also award monetary damages. A court in its discretion has the authority to issue a temporary or permanent injunction to prevent or restrain copyright infringement (Copyright Act, sec. 502).

The copyright owner may recover actual damages suffered as a result of the infringement plus any of the infringer's profits "that are attributable to the infringer and are not taken into account in computing the actual damages" [Copyright Act, sec. 504(b)]. Proving actual damages by the plaintiff based on a copyright infringement requires the plaintiff to show an actual loss. The statute establishes the liability of a copyright infringer for either "the copyright owner's actual damages and any additional profits of the infringer" or statutory damages [sec. 504 (a)]. The plaintiff can also recover any profits that the copyright infringer made by virtue of the infringement. Here the copyright owner need only prove the infringer's gross revenue, and the infringer must prove its deductible expenses and profits not related to the infringement [sec. 504 (b)]. The Copyright Act allows recovery by way of statutory damages, in lieu of actual damages, based on the discretion of the trial judge. Statutory damages range from $750 to $30,000 for each copyrighted work infringed [sec. 504 (c)(1)]. If the court finds that the infringed material was willful, statutory damages can run as high as $150,000 per work infringed [sec. 504 (c) (2)]. In addition, the plaintiff may ask for court costs and reasonable attorney's fees in the discretion of the trial judge. Courts have awarded prejudgment interest on damages for copyright infringement (*Kleier Adver., Inc. v. Premier Pontiac, Inc.,* 1990).

THE DIGITAL MILLENNIUM COPYRIGHT ACT

On October 28, 1998, Title 17 of the Copyright Act was amended by the Digital Millennium Copyright Act (DMCA). Its purpose, in general, is to protect copyright owners from the circumvention of technologies used by them to manage the control and use of the digital content of their copyrighted works and, in certain circumstances, to limit the liability of online service providers for infringement. The broader purpose of the DMCA was to have the copyright act comply with the World Intellectual Property Organization (WIPO) copyright treaty adopted by many countries.

Circumvention of "Digital Locks"

Digitized copyrighted material allows a user to make perfect multiple copies at practically no expense. For example, digitized music, movies, video games, and e-books can be reproduced and sent to countless others in violation of a copyright owner's exclusive statutory rights. Copyright owners may prevent massive reproductions of their protected works by using software products that provide technological locks and controls. The DMCA prohibits a user from decrypting digital locks by circumvention access commonly referred to as the anti-circumvention provision [Copyright Act, sec. 1201 (a)(l)(A)].

Copyright and Management Systems

The DMCA provides that circumvention of a digital lock on a digitized product such as a DVD, e-book, or video game is to "descramble a scrambled work, to decrypt an encrypted work, or otherwise avoid, bypass, remove, deactivate, or impair a technological protection measure" [Copyright Act, sec. 1201 (a)(2000)]. Management systems that use digital locks on their copyrighted works deny access to the digitized product. It is now a violation of the DMCA to circumvent this access control, including the free distribution of software that might provide such anticircumvention.

The DMCA provides copyright owners with the right to use circumvention technology that gives them the ability to manage the use of their copyrighted works with respect to how many copies, if any, can be made of the work, how long the copyrighted digitized work will last, or any other control they may wish to place on the reproduction of the original copyrighted work. As software engineers continue to discover ways of controlling digitized products, there is a growing controversy regarding the balance in the copyright act between the rights of the copyright owner and the "fair use" of others.

Trafficking in Circumvention Tools

The DMCA defines the trafficking in circumvention tools as "the manufacture, import, [or] offer to the public, [of] any technology designed for the purpose of circumventing a technological measure that effectively controls access to a work protected under this title" [Copyright Act, §1201 (2)]. This trafficking provision was addressed in a DVD encryption case, *Universal v. Corley* (2001). In this case, motion picture studios brought a lawsuit under the DMCA to enjoin Internet Web owners from posting computer software that decrypted digitally encrypted movies on DVDs and from posting hyperlinks to Web sites that made decryption software available. The lower court granted an injunction of the online distribution of DeCSS that was a circumvention software product used to decrypt the DVD movie. The appellant court affirmed the judgment and ruled as follows:

> we know of no authority for the proposition that fair use, as protected by the Copyright Act, much less the Constitution, guarantees copying by the optimum method or in the identical format of the original . . . the DMCA does not impose even an arguable limitation on the opportunity to make a variety of traditional fair use of DVD movies, such as commenting on their content, quoting excerpts from their screenplays, and even recording portions of the video images and sounds on film or tape by pointing a camera, a microphone at a monitor as it displays the DVD movie.

This case establishes a prohibition on trafficking in decrypting digital encrypted movies but continues to allow the limited application of fair use of copyrighted material.

Exemptions to the Circumvention Provisions

The DMCA has seven exemptions that, in limited circumstances, allow institutions and individuals to circumvent encryption devices used by copyright owners. They are as follows:

- Libraries, archives and educational institutions [sec. 1201 (d)]
- Law enforcement and intelligence gathering agencies [sec. 1201 (e)]
- Reverse engineering [sec. 1201 (f)]
- Encryption research [sec. 1201 (g)]
- Protecting minors from objectionable material [sec. 1201 (h)]
- Protecting the privacy of personal identifying information [sec. 1201 (i)]
- Security testing [sec. 1201 (j)]

Libraries

If a library were to receive an e-book or a DVD video and wanted to have it reviewed by a number of scholars before it decided to acquire it as part of its collection, would the digital lock on the DVD or e-book prevent it from having the material reviewed simultaneously by various scholars? This exemption has serious limitations on its use. The DMCA allows the exemption only for the limited time it takes the library to make the acquisition decision [17 U.S.C. sec. 1201 (d)(1)(A)]. Decrypting the DVD or other digitized product could be a serious budgetary constraint in that employing cryptographers to engage in this process may, in effect, eliminate this exemption.

Law Enforcement Agencies

This exemption has special application in government antiterrorist activities. Law enforcement agencies can circumvent digital locks in any system that places the government's computer systems at risk.

Reverse Engineering

Assuming that reverse engineering is permitted under copyright law, a software engineer may reverse a copyrighted program to determine whether it can achieve interoperative ability of its own computer program with other programs.

Encryption Research

Encryption research is a discipline of encryption technology in which software engineers and others engage in research to break digital locks. This decryption technology is allowed when the researcher attempted to obtain authorization from the copyright owner before the circumvention of the encrypted program. The DMCA further provides three additional factors to determine whether the court should grant the exemption to the encryption researcher: the encryption research must be "reasonably calculated to advance research on encryption technology in a manner that facilitates infringement" [sec. 1201 (g)(3)(A)], "the researcher must qualify as being trained or experienced in the field of encryption technology" [sec. 1201 (g)(3)(B)], and the encryption researcher must notify the copyright owner of its findings on the research [sec. 1201 (g)(3)(C)]. A serious question to consider and debate is whether this provision has a chilling effect on encryption research because it requires the researcher to notify the copyright owner who may then initiate a claim based on violation of the DMCA.

Protecting Minors From Objectionable Material on the Internet

This provision of the DMCA would allow software products that circumvent the distribution of pornographic sites on the Internet to prevent minors from accessing this objectionable material [17 U.S.C. sec. 1201 (h)].

Privacy Protection

Collection of personal information by a computer user on a device similar to "cookies" used by a Web site may be circumvented under this exemption that has a number of significant limitations. If the Web site posts a privacy policy that provided adequate notice to the user that "cookies" are employed to collect personal identifying information, the exemption would not apply [sec. 1201 (i)(1)(B)]. As stated under the exemption for libraries, consumers could not afford to analyze source code and decrypt a digital protection system such as "cookies" to prevent the dissemination of their personal information.

Security Testing

Testing security systems, such as firewalls, is an important part of electronic commerce. For instance, the security of financial information including credit card information used to purchase a product online must be tested constantly to ensure its security. The DMCA allows circumvention to test security systems [sec. 1201 (j)]. The information acquired from testing this security system must be used exclusively to improve the security and not to infringe the copyrighted product.

Civil and Criminal Remedies for Violating the DMCA

Because the DMCA is a federal statute, a copyright owner alleging infringement must bring a suit in the federal district court. The trial judge has the authority to grant injunctive relief or monetary damages (or both) if there is a violation of the DMCA anticircumvention provisions. The statute allows for the recovery of actual damages as well as reasonable attorney's fees. In the event that there is a willful violation of the anticircumvention provisions of the DMCA used for commercial, private, or financial gain, the court can order significant fines or a criminal action can be prosecuted to imprison the violator.

One can readily see that the DMCA is a statutory system granting new rights to digital copyright owners who attempt to use encryption technology such as "digital locks" to prevent others from copying, viewing more than once, or in various ways attempting to prevent the technological managing of its copyrighted material.

Safe Harbor (Section 512) for Online Service Providers Under the DMCA

In 1998, Congress codified the online copyright Infringement Liability Limitation Act as Section 512 of the Digital Millennium Copyright Act. The purpose behind the safe harbor is to provide to those entities that qualify as service providers, limitations on the remedies available to the copyright owner. The theory is that service providers act only as a conduit for another's infringing material. In certain instances, the service provider should not be held liable for the monetary damages resulting from this infringement.

Online Service Providers

Search engines, Internet service providers, hosting services, and Web sites with multiple links to third parties provide network access to subscribers and customers who may post materials that infringe on copyrights. In this capacity, the customers or subscribers are direct infringers violating the copyright owner's exclusive statutory rights to reproduce the work. The direct infringer is held primarily liable for the infringement.

Copyright law also allows claims against secondary infringers under the tort theories of contributory infringement and vicarious infringement, however. A contributory infringer is a person or entity that has knowledge or had reason to know of the direct infringement and provides a facility for that infringing process. A vicarious infringer need not have knowledge of the direct infringement but had the right to supervise and control the infringers' activities and incurred a financial benefit as a result of the direct infringement. Both contributory infringement and vicarious infringement are referred to as secondary liability.

The federal common law judicial doctrines of the copyright infringement theories of direct and secondary liability become potential copyright claims in view of the function of online service providers. They may be directly liable for copyright infringement by violating the owner's exclusive reproduction rights by intentionally permanently posting copyrighted material on its bulletin board system. If the service provider was properly notified of the infringement and failed to take down the infringing material that, under the circumstances, materially caused or contributed to the infringement, it could be liable for contributory infringement.

Section 512's safe harbor allows avoidance of secondary liability when a service provider posts infringing material on a bulletin board without knowledge, notice, or awareness of its infringing nature and "takes down" the infringing material "expeditiously" after receiving notice of the infringement by the copyright owner [sec. 512 (c)]. The courts generally first determine if contributory liability exists under the federal common law apart from Section 512's safe harbor.

Activities Eligible for Safe Harbor

To qualify for the safe harbor provisions the service provided must engage in any of the following activities:

- Transient holding and transmission of content without modification (so-called store and forward services in which the service provider does not select or review the material)
- Temporary storage of content provided by others ("system caching" in which the service provider temporarily stores someone else's material to provide user's access)
- Storage of material at the user's discretion ("posting" of another's material on a "bulletin board service")
- Linking users to other online locations posting infringing material ("linking" has special significance to a site with multiple links)

Limitation on Remedies Under the Safe Harbor

Section 512 limits the remedies available to the copyright holder. If the defendant fits one of the eligibility categories, Section 512(a) precludes monetary liability and limits injunctive relief to removing the infringing material.

Eligibility for Section 512 "Safe Harbor"

A service provider must comply with the following statutory requirements to be eligible for the safe harbor's limitation on remedies: (a) it must adopt and implement a policy for terminating subscribers and account holders who are repeat infringers and for notifying them of that policy, (b) it must "accommodate and . . . not interfere with standard technical measures" for protecting copyright works [sec. 512 (i)(1)(B)].

Statutory compliance. The limitation on remedies on all of the service providers activities requires further compliance with Section 512. In each case, the service provider cannot have actual notice of the copyright infringement nor can it acquire any financial gain as a result of the infringement. Furthermore, the statute states that the limitation on liability for the online service provider applies only if the service provider has designated an agent to receive notification of the claimed copyright infringement.

Notice on the online service provider's terms of use. A proper notice, generally found on the Web site's terms of use, includes identification of the agent to receive notice of an alleged infringement. This notice is further made accessible to the public by providing to the U.S. Copyright Office the following information: the name, address, phone number, e-mail address of the agent to receive notice, and other information that the registrar of copyright may deem appropriate.

Register of copyright directory of agents. Section 512 of the DMCA further provides that "the registrar of copyrights shall maintain a current directory of agents available to the public for inspection, including through the Internet, in both electronic and hardcopy formats, and payment of a fee by a service provider to cover the cost of maintaining the directory."

Replacement by the online service provider of removed material claimed to be a copyright infringement. Under the DMCA, an online service provider is entitled to notify its customers when their materials have been removed based on a copyright infringement claim [sec. 512 (g)]. The statute provides " upon receipt of a *counter notification*, the online service provider may replace the removed material and notify the copyright owner, wherein the owner has 14 days to file an action in the Federal district court for copyright infringement against the direct infringer."

Contents of the counter-notification. The contents of the counter-notification must include substantially the following material: (a) a physical or electronic signature of the subscriber, (b) identification of the material that has been removed or to which access has been disabled, (c) the location at which the material appeared before it was removed, (d) a good-faith belief that the material was removed as a result of mistake, (e) the subscriber's name, (f) misidentification of the material, (g) subscriber's address and telephone number, and (h) a statement that the subscriber's consent to the jurisdiction of the federal district court for the judicial district in which the address is located.

Limitation on the liability of nonprofit educational institutions. Under employment law, an employer is liable for the wrongdoing of its employees causing injury to others while carrying on employment business. An exception is made to this doctrine when a public or other nonprofit organization of higher education is an online service provider. The statute provides the following under Section 512 (e)(1): "when a faculty member or graduate student who is an employee of such institution is performing a teaching or research function, such faculty member or graduate student shall be considered to be a person other than the institution and their knowledge or awareness of the infringing activity shall not be attributed to the institution." This exemption from liability of the faculty member or graduate student is available providing the educational institution has not, within the preceding three-year period of notice of the infringement, received more than two notifications of claimed infringement by that faculty member or graduate student and that the educational institution provides to all users of its online system compliance with the laws of the United States relating to copyright. Educational institutions must have copyright law policy that is adequately published requiring all faculty members to immediately notify the institution of any alleged infringements based upon material that may be posted on their Web sites. Section 512 of the DMCA presents a public policy position that service providers of such essential functions on the Internet as search engines, bulletin board services, Web sites with multiple links, educational institutions, and others that may qualify as an online service provider are granted federal immunity from contributory or vicarious infringement based on their relationships with their customers or subscribers who may be direct infringers. Strict compliance with the DMCA is necessary for the online service provider to have the benefit of this safe harbor provision.

INTERNATIONAL COPYRIGHT LAWS

Because the Internet is an international medium, courts around the world hear copyright infringement cases. The World Trade Organization (WTO) comprises a number of countries, including the United States. It observes the provisions of intellectual property agreements signed by the member countries. Two of the most important agreements are the Berne Convention for the Protection of Literary and Artistic Works (Berne Convention) and the Trade-Related Aspects of Intellectual Rights.

WTO member countries agree to a dispute-resolution policy established by WTO that resolves intellectual property copyright disputes between its member countries. The Berne Convention mandates its members to protect copyrighted works, and it established four principles of membership compliance. These include similar national treatment for all member countries, nonconditional protection so that no formalities are required to protect a copyrighted work (such as the use of the copyright symbol, ©). Although no longer required by the Berne Convention, use of the symbol remains good practice to disclose the intent of preserving the copyrighted material by including on documents, including a Web site, the copyright symbol followed by the year and the name of the entity claiming ownership in the work. This has the secondary effect of making it difficult for an infringer to claim that this was an innocent infringement and argue for a consequent reduction in damages. The third principle provided by the Berne Convention provides that there is copyright protection independent of the country of origin so that registration in one member country protects in all member countries. The fourth principle provides for common rules that establish procedures for granting copyright protection.

The World Intellectual Property Organization Copyright Treaty has extended the Berne Convention's copyright protection to include computer software (see http://wipo.int; accessed August 18, 2002). Businesses that intend to market their product internationally should be aware that there is no worldwide automatic protection of copyright. National laws of every country and various treaties that a country may have entered regarding copyright protection should be explored through appropriate legal counsel (Copyright Circular 38A; available at http://lcweb.loc.gov/copyright; accessed August 18, 2002).

CONCLUSION

Digital content on Web pages that are original and creative are appropriate subject matter for copyright protection. Although registration is not necessary to create a copyright, upon federal registration in the Copyright Office, the copyright author, who may be a company or an individual, is granted statutory benefits including exclusive statutory rights over the content of the Web pages, access to the federal court for an infringement suit, statutory damages in lieu of actual damages, and a presumption of copyright ownership. The benefits of copyright registration in the Copyright Office should always be discussed with legal counsel.

If an e-business Web site is interactive through the use of a bulletin board service, a search engine, or multiple links, the site may qualify as a service provider under the Digital Millennium Copyright Act and be granted the benefits of "safe harbor," which limits the remedies available to the copyright owner. Vigilant copyright management should be an important function of the e-business to protect its valuable intellectual copyright property. Management should work closely with legal counsel to constantly ensure statutory compliance with the Copyright Act and, when applicable, foreign copyright laws.

GLOSSARY

Circumvention under the DMCA Bypassing a technological device used by the copyright owner to prevent copying of the work.

Contributory copyright infringement A person who knew, or should have known, of the direct infringement and who thus induces, causes, or materially contributes to the infringing conduct of another.

Copyright Act Federal statutory law that grants to authors of original works, for a limited time, exclusive statutory rights over the copyrighted works.

Digital Millennium Copyright Act (DMCA) A federal statute that generally prohibits circumvention of any device used by the copyright owner to prevent copying of the work; it further provides a "safe harbor" for online service providers who show copyrighted works on its Web site.

Direct copyright infringement The entity that performs the copyright infringement.

Exclusive rights of copyright owner Statutory rights, under the Copyright Act, granted to the author of original works for a limited time.

Fair use A common law doctrine now part of the Copyright Act that allows limited use of the copyright material if it is educational, a small part of the copyright work, and does not affect the potential or actual value of the work.

First sale doctrine Right of the purchaser of a copyright work to dispose of it in any way, including selling it.

Independent contractor A person hired to work on a project without the supervision and control of another.

Information location tool Under the DMCA, a directory, index, reference, pointer, or multiple hypertext link.

Online service provider An entity offering connections for digital online communications.

Primary infringement liability *See* direct infringement.

"Safe Harbor" provision of the DMCA A process to limit the remedies of a copyright owner brought against an online service provider.

Secondary infringement liability Contributory infringement or vicarious infringement, providing there is evidence of a direct infringement.

Sonny Bono Copyright Term Extension Act (1998) Federal law that extended the duration of a copyright by 20 years.

Theories of copyright infringement The plaintiff in a copyright case must prove ownership of the copyright, violation of an exclusive right, and a theory of copyright infringement that must include a direct infringement and may include a contributory or a vicarious infringer.

Vicarious copyright infringement An entity that has the right and ability to supervise the direst infringement and had a financial interest in the infringed copyrighted work.

Work for hire Copyrighted work authored by an employee, as part of an employment contract, that is owned by the employer.

World Trade Organization An international group of nations (including the United States) that overviews the provisions of intellectual property signed by member countries.

CROSS REFERENCES

See *Patent Law; Trademark Law.*

REFERENCES

A&M Records, Inc., v. Napster, Inc., 239 F.3d 1004, 1022 (9th Cir. 2001).

Bateman v. Mnemonic, Inc., 79 F.3d 1532, 1542 n. 22 (11th Cir., 1995).

CCC Information Services, Inc., v. McAllen Hunter Market Reports, Inc., 44 F.3d. 612 (2nd Cir. 1994).

Community for Creative Nonviolence v. Reid, 440 U.S. 739 (1989).

Computer Assoc. Int'l. v. Altai, Inc., 982 F.2d 693, 696 (2d Cir. 1992).

Copyright Act, 17 U.S.C. 101–1205.

Data Gen. v. Grumman Sys. Support Corp., 36 F.3rd 1147, 1160 (1st Cir. 1994)

Dratler, J. (2002). *Cyberlaw: Intellectual property in the digital millennium.* New York: Law Journal Press.

eBay, Inc. v. Bidder's Edge, Inc., 100 F. Supp. 2d 1058 (N.D. Cal. 2000).

Eldred v. Ashcroft, 122 S. Ct. 1170 (2003).

Feist Publications, Inc., v. Rural Telephone Services Inc., 499 U.S. 340 (1991).

Gershwin Publ'g Corp. v Columbia Artist Mgmt., Inc., 443 F. 2d 1159, 1162 (2nd Cir. 1971).

Intellectual Reserve, Inc. v. Utah Lighthouse Ministries, Inc., 75 F. Supp. 2d 1240 (D. Utah l999).

Irwin, B., Koenigsberg, I. F., and Spelman, K. (2000). *Understanding basic copyright law*. New York: Practicing Law Institute.

Joyce, C., W., Leaffer, M., and Jaszi, P. (2000). *Copyright law* (5th Ed.). New York: Matthew Bender & Company, Inc.

Keller, B. and Cunard, J. (2002). *Copyright law, a practioner's guide*. New York: Practicing Law Institute.

Kelly v. Arriba Soft Corp., 280 F.3d 934 (9th Cir. 2002).

Kleier Adver., Inc. v. Premier Pontiac, Inc., 921 F.2d 1036, 1040–42 (10th Cir. 1990).

M.C.B. Homes, Inc. v. Ameron Homes, Inc., 903 F.2d 1486, 1488 & n.4 (11th Cir. 1990).

New Era Publ'ns Int'l, ApS v. Henry Holt & Co., Inc., 873 F.2d 576, 584 (2nd Cir. 1989).

Nimmer, M. and Nimmer, D. (2002)-Nimmer on Copyright. New York: Matthew Bender & Company, Inc.

ProCD, Inc. v. Zeidenberg, 86 F.3d 1447 (7th Cir. 1996).

S&H Computer Sys. v. SAS Inst., Inc., 568 F. Supp. 426, 422 (1983).

Sony Corporation of America v. Universal City Studios, Inc., 464 U.S. 417 (1984).

Step-Saver Data Sys., Inc. v. Wyse Tech., Inc., 939 F.2d 91 (3d Cir. 1991).

Steward v. Abend, 495 U.S. 207, 236 (1990).

Suntrust Bank v. Houghton Mifflin Co., 286 F.3d 1257, 1261 (11th Cir. 2001).

Synercom Technology, Inc. v. University Computing Co., 4622 F. Supp. 1002, 1013 (1978).

Universal v. Corley, 273 F.3rd 429 (2nd Cir. 2001).

Universal City, Inc. v. Reinerdes, 111 F. Supp. 2d 294, 321–22.

FURTHER READING

Copyright Act. Retrieved August 12, 2002, from http://www4.law.cornell.edu/uscode/17/

Copyright Basics, Circular 1. Retrieved August 18, 2002, from http://www.loc.gov/copyright/circs/circ1.html

Digital Millennium Copyright Act, legislative history. Retrieved August 12, 2002, from http://www.lib.umich.edu/copyright/dmca.html

Explanations of online copyright issues. Retrieved August 12, 2002, from http://www.benedict.com

U.S. Copyright Office. Retrieved August 12, 2002, from http://lcweb.loc.gov/copyright

World Intellectual Property Organization. Retrieved August 12, 2002, from http://www.wipo.int (Date of access: August 12, 2002)

Customer Relationship Management on the Web

Russell S. Winer, *New York University*

INTRODUCTION

The essence of the information technology (IT) revolution and, in particular, the World Wide Web is the opportunity afforded companies to improve their interactions with their customers. The Web allows companies to build better relationships with customers than was previously possible in the offline world. By combining the abilities to respond directly to customer requests and to provide the customer with a highly interactive, customized experience, companies have a greater ability today to establish, nurture, and sustain long-term customer relationships than ever before. These online capabilities complement personal interactions provided through salespeople, customer service representatives, and call centers. At the same time, companies can choose to exploit the low cost of Web customer service to reduce their service costs and offer lower quality service by permitting only electronic contact. The flexibility of Web-based interactions thus permits firms to choose to whom they wish to offer services and at what quality level.

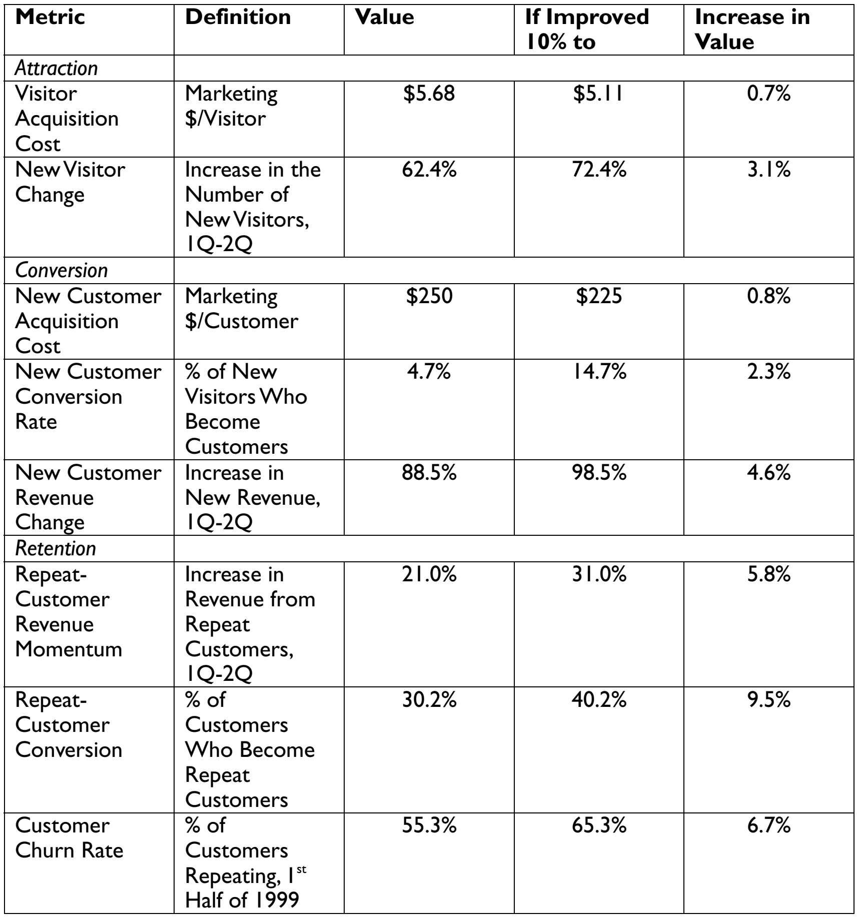
Indeed, this revolution in customer relationship management (CRM) has been referred to as the new “mantra” of marketing. (If all of the components of the CRM system are Web-based, the term “eCRM” (electronic CRM) is sometimes used.) Companies such as Siebel, E.piphany, Oracle, Broadvision, Net Perceptions, Kana, Salesforce.com, and others have developed CRM products that do everything from track customer behavior on the Web, to predicting their future moves, to sending direct e-mail communications. Although the IT spending slowdown of 2000–2002 has put a crimp in the industry’s sales, Morgan Stanley Dean Witter & Co. estimates the total potential for CRM software to be more than $48 billion (Enrado, 2002).

The need to better understand customer behavior and the interest of many managers to focus on those customers who can deliver long-term profits has changed how marketers view the world. Traditionally, marketers have been trained to acquire customers, either new ones who have not bought the product before or those who are currently competitors’ customers. This has required heavy doses of mass advertising and price-oriented promotions to customers and channel members. Although the revolution in improved customer service programs and measurement since the 1980s brought a welcome emphasis on the customer, today, particularly for the company’s “best” customers, the tone of the conversation has changed from customer acquisition to retention. This requires a different mindset and a new set of tools. A good thought experiment for an executive audience is to ask them how much they spend or focus on acquisition versus retention activities. It is difficult to perfectly distinguish the two activities from each other, but the answer is usually that acquisition dominates retention.

The impetus for this interest in CRM came from Reichheld (1996), who demonstrated dramatic increase in profits from small increases in customer retention rates and popularized the notion of the lifetime value of customers. His studies showed that as little as a 5% increase in retention had impacts as high as 95% on the net present value delivered by customers. Other studies done by consultants such as McKinsey have shown that repeat customers generate more than twice as much gross income as new customers. The considerable improvements in technology and innovation in CRM-related products have made it much easier to deliver on the promise of greater profitability from reduced customer “churn.”

For example, Figure 1 shows the results from a 1999 McKinsey study on the simulated impact of improvements in a number of customer-based metrics on the market value of Internet companies. The metrics are divided into three categories: customer attraction, customer conversion, and customer retention. As can be seen, the greatest leverage comes from investments in retention. If revenues from repeat customers, the percentage of customers who repeat purchase, and the customer churn rate each improves by 10%, the company value was found to increase (theoretically) by 5.8%, 9.5%, and 6.7%, respectively.

A problem is that although managers understand what the letters CRM stand for, the interpretation of the concept can vary by person. For some companies, the main CRM application is a telephone call center. For others, it is mass customization or developing products that fit individual customers’ needs. For salespeople, it means sophisticated customer contact software. For IT consultants, CRM translates into complicated technical jargon related



Metric	Definition	Value	If Improved 10% to	Increase in Value
Attraction				
Visitor Acquisition Cost	Marketing $/Visitor	$5.68	$5.11	0.7%
New Visitor Change	Increase in the Number of New Visitors, 1Q-2Q	62.4%	72.4%	3.1%
Conversion				
New Customer Acquisition Cost	Marketing $/Customer	$250	$225	0.8%
New Customer Conversion Rate	% of New Visitors Who Become Customers	4.7%	14.7%	2.3%
New Customer Revenue Change	Increase in New Revenue, 1Q-2Q	88.5%	98.5%	4.6%
Retention				
Repeat-Customer Revenue Momentum	Increase in Revenue from Repeat Customers, 1Q-2Q	21.0%	31.0%	5.8%
Repeat-Customer Conversion	% of Customers Who Become Repeat Customers	30.2%	40.2%	9.5%
Customer Churn Rate	% of Customers Repeating, 1st Half of 1999	55.3%	65.3%	6.7%

Figure 1: Impact of a 10% improvement in indicator on the current value of e-commerce firms. Source: Cigliano, Georgiadia, Pleasance, & Whalley (2000).

to terms such as OLAP (online analytical processing) and CICs (customer interaction centers).

One way to better understand CRM is to decompose it term by term:

1. *Customer.* Clearly, an important focus of CRM is on customers,but which customers? Although it is most natural to consider customers as being those parties that purchase a company's services or products, Siebel has developed a version of its software for internal, human resource management use for retaining valued employees.
2. *Relationship.* Marketers have studied brand loyalty for nearly 50 years. Thus, the notions that repeat purchasing is important and loyal customers are valuable are not new. The key to CRM, however, is understanding that different customers want different kinds of relationships—some want to single source products, some want multiple vendors. The idea of CRM is to customize the relationship to what benefits and keeps the customer.
3. *Management.* The point of the "M" is that relationships have to be managed; they do not survive without the active participation of the company seeking to extend them.

To better implement CRM within organizations, I have developed a series of steps that managers should follow in implementing a CRM program. What do managers need to know about their customers, and how is that information used to develop a complete CRM perspective? Figure 2 shows a basic model, which contains a set of seven basic components:

1. A database of customer activity;
2. Analyses of the database;
3. In light of these analyses, decisions about which customers to target;
4. Tools for targeting the customers;
5. Knowledge of how to build relationships with the targeted customers;
6. Awareness of privacy issues; and
7. Metrics for measuring the success of the CRM program.

Note that these steps are generic in nature, that is, they apply whether the context is Web-based or non-Web-based CRM. The basics are the same in either context. The Web has made the goal of improved customer satisfaction and retention easier to reach, however, because of the increased interactivity and personalization that the medium affords. Taken together, these two characteristics

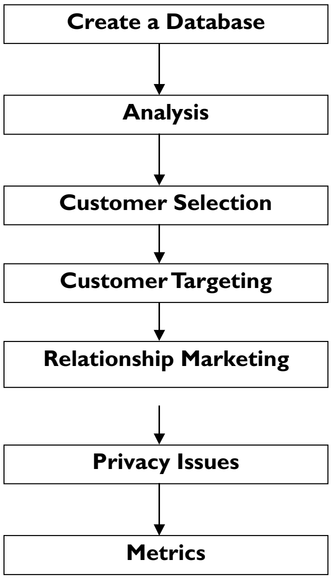


Figure 2: Customer relationship management model.

of the Web have made customer retention activities resonate with customers as they are tailored to them. The purpose of this chapter is to provide the reader with an understanding about how interactivity and personalization have dramatically improved the generic CRM system.

CREATING A CUSTOMER DATABASE

A necessary first step to a complete CRM solution is the construction of a customer database or information file (Glazer, 1999). This is the foundation for any CRM activity. For Web-based businesses, constructing a database should be a relatively straightforward task as the customer transaction and contact information is accumulated as a natural part of the interaction with customers. For existing companies that have not previously collected much customer information and are moving to the Web, the task will involve merging historical customer contact data from internal sources such as accounting and customer service with the data provided by the Web-based interactions.

What should be collected for the database? Ideally, the database should contain information about the following:

- *Transactions*—This should include a complete purchase history with accompanying details (price paid, SKU, delivery date).
- *Customer contacts*—Today there is an increasing number of customer contact points from multiple channels and contexts. This should include not only sales calls and service requests, but any customer- or company-initiated contact.
- *Descriptive information*—This is for segmentation and other data analysis purposes. Useful information here would include address and demographic information for consumers and location and company information for business customers.
- *Responses to marketing stimuli*—This part of the information file should contain whether the customer responded to a direct marketing initiative, a sales contact, or any other direct contact.
- *The value of the customer*—Many companies calculate the value of each customer, both the current and the long-term or lifetime value. This provides a valuable basis for making resource allocation decisions at the customer level.

The data should also be represented over time.

Companies have traditionally used a variety of methods to construct their databases. Durable goods manufacturers use information from warranty cards for basic descriptive information. Unfortunately, response rates to warranty cards are in the 20–30% range leaving big gaps in the databases. Service businesses are normally in better shape because the nature of the product involves the kind of customer-company interaction that naturally leads to better data collection. For example, banks have been in the forefront of CRM activities for a number of years. Telecom-related industries (long distance, wireless, cable services) similarly have a large amount of customer information. Catalog companies (e.g., Lands' End) also have excellent customer information files.

The following are illustrations of some corporate database-building efforts:

- The networking company 3Com created a worldwide customer database from 50 "legacy" databases scattered throughout their global operations. They built customer records from e-mails, direct mail, telemarketing, and other customer contacts, with descriptive information by department, division, and location.
- Thomson Holidays, the British tour company, developed a Preferred Agents Scheme to enlist the assistance of travel agents in building the database. They collect customer descriptive information and data on trips taken. This enables them to calculate the profit on a per-customer-trip basis.
- Taylor Made, the golf equipment manufacturer, has a database of more than 1.5 million golfers with their names, addresses, e-mail addresses, birthdays, types of courses played, and vacations taken.

These databases are generally made available to all relevant managers through an intranet or some other Web-based interface. This permits managers anywhere in the world to access and add to the customer information.

Companies such as Procter & Gamble and Unilever that sell frequently purchased consumer products have greater problems constructing databases because of a lack of systematic information about their millions of customers and the fact that they use intermediaries (i.e., supermarkets, drugstores) that prohibit direct contact. The challenge is to create opportunities for customer interaction and, therefore, data collection. This can be from running contests to encouraging customer visits to Web sites.

		Customer Interaction	
		Direct	Indirect
Interaction Frequency	High	Banks Telecom Retail	Airlines Packaged Goods Drugs
	Low	Personal Computers Internet Infrastructure	Furniture Autos

Figure 3: Getting more customer interaction. Courtesy of Professor Florian Zettelmeyer of the University of California at Berkeley.

Figure 3 gives a general framework for considering the problems in database construction. Firms in the upper-left-hand quadrant have many direct customer interactions (banks, retail) and therefore have a relatively easy job constructing a database. Firms in the lower-right-hand quadrant have the most difficult job because they interact less frequently with customers and those interactions are indirect (through channels) in nature. Auto and furniture manufacturers are examples here. The other two boxes represent intermediate situations.

The point of this framework is that unless you are in the high-direct box, you have to work harder to build a database. The Thomson Holidays example provides a good illustration of company that uses channel incentives to take a low-frequency product and still obtain customer information. Kellogg has developed a creative solution to the problem through its "Eat and Earn" program where children find a 15-digit code inside cereal boxes and then go to the company's Web site, enter some personal information, and become eligible for free toys. The task for companies is then to move toward the upper-left-hand quadrant through increased customer contact and "event" marketing.

ANALYZING THE DATABASE

Traditionally, customer databases have been analyzed with the intent to define customer segments. A variety of multivariate statistical methods such as cluster and discriminant analysis have been used to group together customers with similar behavioral patterns and descriptive data which are then used to develop different product offerings or direct marketing campaigns (see, for example, Wedel and Kamakura, 1999). Direct marketers have used such techniques for many years. Their goals are to target the most profitable prospects for catalog mailings and to tailor the catalogs to different groups.

More recently, such segmentation approaches have been criticized for the assumption that segments are even reasonably homogeneous (Peppers & Rogers, 1993). Taking a large number of customers and forming groups or segments presumes a marketing effort toward an "average" customer in the group. Given the range of marketing tools available that can reach customers one at a time using tailored messages designed for small groups of customers (what has been referred to as "1-to-1" marketing), there is less need to consider the usual market segmentation schemes that contain large groups of customers (e.g., women 18–24 years of age). Rather, there is increased attention being paid to understanding each "row" of the database—that is, understanding each customer and what he or she can deliver to the company in terms of profits and then, depending on the nature of the product or service, addressing either customers individually or in small clusters.

As a result, although we are not at the point where we can say that segmentation is a dead concept, a new term, "lifetime customer value" (LCV), has been introduced into the lexicon of marketers. The idea is that each row (customer) of the database should be analyzed in terms of current and future profitability to the firm. When a profit figure can be assigned to each customer, the marketing manager can then decide which customers to target.

Although there are a number of formulas that can be used to estimate LCV, one way to start is with a relatively simple formula that uses only the available purchase information in the CIF to calculate each customer's cumulative profitability to the present time:

$$\text{Customer profitability} = \sum_{t} \left[\sum_{j} (\mathrm{P_j}-\mathrm{C_j}) - \sum_{\mathrm{k}} \mathrm{MC_k} \right]$$

where t = the number of past and current time periods measured, j = the number of products purchased in a time period, k = the number of marketing tools used in a time period, P = price, c = cost, and mc = cost of marketing tool (e.g., direct mail).

The formula computes the total profits generated by a customer by taking the total margin generated by the customer (P – C) in each time period from all products and services purchased, subtracting the traceable marketing costs attributable to each customer and then summing over all time periods in the database.

This formula is useful for purposes other than computing customer profitability. An examination of its components shows the levers for increasing individual customer profitability. Profits can be increased by

- Increasing P and j by cross-selling (purchasing more products) or up-selling (purchasing more expensive products),
- Reducing the marketing costs MC over time as the customer loyalty becomes better established, and
- Increasing the number of time periods t that the customer purchases.

To compute the LCV, one must project the customer's generated margins and marketing costs into the future and discount back. This requires a number of assumptions about the nature of a customer's purchasing pattern in the future. A back-of-the-envelope approach to calculating LCV is a margin "multiple" that can be used to multiply the current margin generated by each customer to estimate the LCV (Gupta & Lehmann, 2003). This multiple is

$$R/(1 + I–R).$$

In this formula, R is the retention (loyalty) rate for the product and I is the discount or cost of capital rate used by the company. Thus, for a product with a retention rate of 70% and a discount rate of 12%, one takes the margin generated by each customer and multiplies it by 1.67. This approximates the LCV for that customer.

Other kinds of data analyses besides LCV are appropriate for CRM purposes. Marketers are interested in what products are often purchased together, frequently referred to as market basket analysis. Complementary products can then be displayed on the same page on a Web site. Internet-based firms such as Amazon use collaborative filtering mechanisms that match a customer's past buying habits with similar customers to make product recommendations (Ansari, Essegaier, & Kohli, 2000).

As noted, a new kind of analysis born from the Internet is the clickstream analysis. In this kind of data analysis, patterns of mouse "clicks" are examined from cyberstore visits and purchases to better understand and predict customer behavior (see, for example, Moe & Fader, 2001). The goals are to increase "conversion" rates, the percentage of browsing customers to actual buyers, and sales per customer. For example, companies such as Blue Martini and Net Perceptions sell software that enables Web-based stores to customize their sites in real time depending on the type of customers visiting—that is, their previous buying patterns, other sites visited during the current session, and their search pattern in the cyberstore. An advantage of the Internet is that customer movements on the Web can be tracked, stored, and analyzed in real time to form the basis of increasingly customizable offerings. Brooks Brothers, the retailer of classic men's and women's clothing, is testing a new approach to increasing it e-commerce sales. When a visitor comes to http://www.brooksbrothers.com, software scans the visitor's computer for its Internet (IP, or Internet protocol) address and the user's list of previously visited sites. It then develops a profile of the visitor based on trying to match the information with others in a database of 100 million people. The visitors are then sorted in real time by the site's best guess on gender, marital status, and geography. Subsequently, the visitor is presented with a Web page based on that guess. As a result, a (hypothesized) single male would receive a different page than a woman and an urban male would receive a different product assortment than a rural male.

CUSTOMER SELECTION

Given the construction and analysis of the customer information contained in the database, the next step is to consider which customers to target with the firm's Internet-based and other direct marketing programs. The results from the analysis can be of various types. If segmentation analyses are performed on purchasing or related behavior, the customers in the most desired segments (e.g., highest purchasing rates, greatest brand loyalty) would normally be selected first for retention programs. Other segments can also be chosen depending on additional factors. For example, for promotions delivered by targeted e-mail, if the customers in the heaviest purchasing segment already buy at a rate that implies further purchasing is unlikely, a second tier with more potential would also be attractive. The descriptor variables for these segments (e.g., age, industry type) provide information for deploying the marketing tools. In addition, these variables can be matched with commercially available databases of names to find additional customers matching the profiles of those chosen from the database.

For example, Rust, Lemon and Zeithaml (2000) divide the customer base into Platinum, Gold, Iron, and Lead with decreasing levels of profitability and increasing price sensitivity. The challenge is to develop profiles that clearly distinguish between the segments of the customer "pyramid" so that the marketing manager can develop communications programs to reach them. If individual customer-based profitability is also available through LCV or similar analysis, it would seem to be a simple task to determine on which customers to focus. The marketing manager can use a number of criteria such as simply choosing those customers that are profitable (or projected to be) or imposing an ROI hurdle. The goal is to use the customer profitability analysis to separate customers who will provide the most long-term profits from those who are currently hurting profits. This allows the manager to identify customers who are too costly to serve relative to the revenues being produced and then develop programs that allow them to either become profitable or switch to another supplier. Although this may seem contrary to being customer-oriented, the basis of the time-honored "marketing concept," in fact, there is nothing that says that marketing and profits are contradictions in terms. The 80/20 rule of thumb often holds in approximation: Most of a company's profits are derived from a small percentage of their customers. For example,

- AT&T offers different levels of customer service depending on a customer's profitability in their long-distance telephone business. For highly profitable customers, they offer personalized service. For less profitable customers, you get automated, menu-driven service.
- The wireless provider PageNet raised monthly rates for unprofitable subscribers. Clearly, the intent was to either make them profitable or for them to leave for another company.
- Similarly, Federal Express raised shipping rates for residential customers in expensive-to-serve areas where their volume did not justify normal rates.

The point is that without understanding customer profitability, these kinds of decisions cannot be made.

On what basis should these customer selection decisions be made? One approach would be to take the current profitability based on the above equation. An obvious problem is that by not accounting for a customer's possible growth in purchasing, you could be eliminating a potentially important customer. Customers with high LCV could be chosen, because this does a better job incorporating potential purchases. These customers are difficult to predict, however, and one might include a large number of unprofitable customers in the selected group. No matter what criterion is employed, de-selected customers need to be chosen with care. Once driven away or ignored, unhappy customers can spread negative word-of-mouth quickly, particularly in today's Internet age.

TARGETING THE CUSTOMERS

Mass marketing approaches, such as television, radio, or print advertising, are useful for generating awareness and achieving other communications objectives, but they are poorly suited for CRM because of their impersonal nature. More conventional approaches for targeting selected customers include a portfolio of direct marketing methods such as telemarketing, direct mail, and, when the nature of the product is suitable, direct sales. Writers such as Peppers and Rogers (1993) urged companies to begin to dialogue with their customers through these targeted approaches rather than talking "at" customers with mass media.

In particular, the new mantra, "1-to-1" marketing, has come to mean using the Internet to facilitate individual relationship building with customers. An extremely popular form of Internet-based direct marketing is the use of personalized e-mails. When this form of direct marketing first appeared, customers considered it no different from the "junk" mail that they receive at home and treated it as such, with quick hits on the keyboard's delete button. Sparked by Godin's (1999) call for "permission" based programs whereby customers must first "opt-in" or agree to receive messages from a company, however, direct e-mail has become a popular and effective method for targeting customers for CRM purposes.Companies such as Kana and Digital Impact can send sophisticated e-mails including video, audio, and Web pages. Targeted e-mails have become so popular that an estimated 430 billion e-mail advertisements will be sent in 2002 (O'Connell, 2002).

A study by Forrester Research shows why this is so (Nail, 2000). Figure 4 demonstrates that e-mail is a cost-effective approach to customer retention. Through lower cost per 1,000 names by using the company's own database (the "house" list) and greater clickthrough rates than those afforded by banner advertisements and e-mails sent to lists rented from suppliers, companies can reduce their cost per sale dramatically.

Some examples are the following:

- Southwest Airlines' e-mail-based Click 'n Save program has 2.7 million subscribers. Every Tuesday, the airline sends out e-mails to this database of loyal users containing special fare offers.
- The bookseller Borders (Borders.com, Borders and Waldenbooks offline retailers) collected all of its customer information into a single database. The company then uses e-mails tailored to customers' reading interests to alert them about upcoming releases.
- The Phoenix Suns basketball team sends streaming video messages from its players promoting new

	Customer Acquisition			Customer Retention	
	Direct Mail to Rented list	Banner Advertising	E-mail to Rented List	Direct Mail to "House" List	E-mail to "House" List
Cost Per 1,000	$850	$16	$200	$686	$5
Click-Through Rate	N/A	0.8%	3.5%	N/A	10%
Purchase Rate	1.2%	2.0%	2.0%	3.9%	2.5%
Cost Per Sale	$71	$100	$286	$18	$2

Figure 4: E-mail generates the lowest retention costs. Source: Nail, J., 2000.

ticket packages and pointing them to the team's Web site.

The problem with targeted e-mails today is that many companies are not using the permission-based approach that Godin advocated, resulting in consumer dissatisfaction with increased "spam," because perhaps as many as one third of all the e-mail messages that individuals receive are uninvited. This has reduced the clickthrough rates on rented lists from the 3.5% shown in Figure 4 to about 1.8% in 2002.

RELATIONSHIP PROGRAMS

The overall goal of relationship programs is to deliver a higher level of customer satisfaction than competing firms deliver. There has been a large volume of research in this area (see, for example, Oliver, 1997, and Zeithaml & Bitner, 2000). From this research, managers today realize that customers match realizations and expectations of product performance and that it is critical for them to deliver such performance at higher and higher levels as expectations increase because of competition, marketing communications, and changing customer needs. In addition, research has shown that there is a strong, positive relationship between customer satisfaction and profits (Anderson, Fornell, & Lehmann, 1994). Thus, managers must constantly measure satisfaction levels and develop programs that help to deliver performance beyond targeted customer expectations.

Although customer contact through direct e-mail offerings is a useful component of CRM, it is more of a technique for implementing CRM than a program itself. Relationships are not built and sustained with direct e-mails themselves but rather through the types of programs that are available for which e-mail may be a delivery mechanism.

A comprehensive set of relationship programs is shown in Figure 5 and includes customer service, frequency and loyalty programs, customization, rewards programs, and community building.

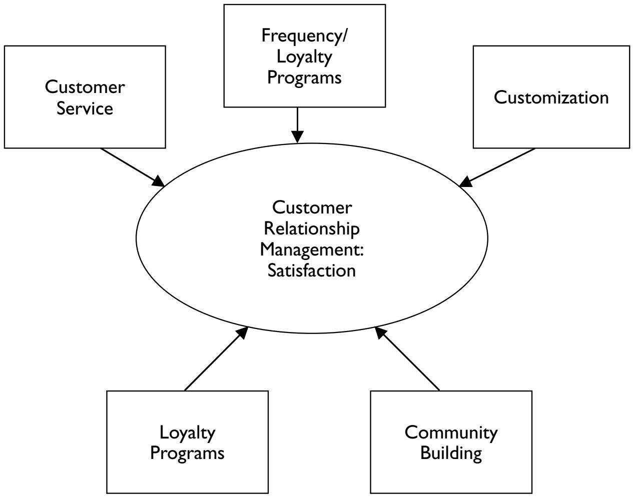

Customer Service
Frequency/ Loyalty Programs
Customization
Customer Relationship Management: Satisfaction
Loyalty Programs
Community Building

Figure 5: Customer retention programs.

Customer Service

Because customers have more choices today and the targeted customers are most valuable to the company, customer service must receive a high priority within the company. In a general sense, any contact or "touch points" that a customer has with a firm is a customer service encounter and has the potential either to gain repeat business and help CRM or to have the opposite effect. Programs designed to enhance customer service are normally of two types. *Reactive* service is where the customer has a problem (product failure, question about a bill, product return) and contacts the company to solve it. Most companies today have established infrastructures to deal with reactive service situations through 800 telephone numbers, faxback systems (automated systems by which customers request information from a menu, and the response is faxed back), e-mail addresses, and a variety of other solutions. *Proactive* service is a different matter; this is a situation in which the manager has decided not to wait for customers to contact the firm but rather to be aggressive in establishing a dialogue with customers before complaining or other behavior sparks a reactive solution. This is more a matter of good account management where the sales force or other people dealing with specific customers are trained to reach out and anticipate customers' needs.

A variety of systems leveraging the Web assist both kinds of service. Charles Schwab has established MySchwab, which allows customers to create personal Web pages linking them to all Schwab services, including stock quotes, trading, and retirement planning analyses. In this way, the company empowers customers to deliver their own service. Other Web-based services such as LivePerson and netCustomer are software products that, when added to a company's Web site, provide customers with the ability to interact with service representatives in real time. Companies such as Kmart are investing large amounts of money into kiosks that provide information on product availability, order status, and a variety of other service-related topics.

Loyalty and Frequency Programs

Loyalty programs (also called frequency programs) provide rewards to customers for repeat purchasing. A recent McKinsey study (Cigliano, Georgiadis, Pleasance, & Whalley, 2000) found that about half of the 10 largest U.S. retailers in each of the top seven sectors (category killers such as Home Depot, department stores, drugstores, gasoline, grocery, mass merchandisers, specialty apparel) have such programs, and there have been similar findings in the United Kingdom. The McKinsey study also identified the three leading problems with these programs: they are expensive, mistakes can be difficult to correct because customers see the company as taking away benefits, and, perhaps most important, there are large questions about whether they work to increase loyalty or average spending behavior (Dowling & Uncles, 1997; Reinartz & Kumar, 2000). A problem that can be added to this list is that because of the ubiquity of these programs, it is increasingly difficult to gain competitive advantage. As the managers for the airlines will attest, however, loyalty programs can be successful by increasing customer switching costs and building barriers to entry. In addition, in some industries, such programs have become a competitive necessity.

Some Web-based companies provide incentives for repeat visits to Web sites, the most notable being MyPoints.com. Although these have not been wildly successful (a number of them crashed and burned in the Internet implosion), it is clear that the price orientation of many Web shoppers creates the need for programs that can generate loyal behavior.

Customization

The notion of mass customization goes beyond 1-to-1 marketing because it implies the creation of products and services for individual customers, not simply communicating with them. Dell Computer popularized the concept with its build-to-order Web site. Other companies such as Levi Strauss, Nike, and Mattel have developed processes and systems for creating customized products according to customers' tastes. Slywotzky (2000) referred to this process as a "choiceboard" by which customers take a list of product attributes and determine which they want. The idea is that it has turned customers into product *makers* rather than simply product *takers*.

For example, a visit to the Levi Strauss Original Spin Web page (http://www.levi.com/Original_Spin/) vividly demonstrates the choiceboard concept. After providing size information at an Original Spin retail store, the customer has a choice of 7 models of jeans (e.g., classic, low cut), 8 leg openings (e.g., tapered, straight), 13 fabrics, and 2 flys (button, zipper). This gives the customer 1,456 options with a perfect fit.

Shapiro and Varian (1999) argued that customization or "versioning" is cheap and easy to do with information goods. In this instance, although a different product would not be built for each individual customer as in the Levi's case, it is relatively easy to develop a number of product configurations for various market segments based on their needs. It is, of course, easier to do this for services and intangible information goods than for products, but the examples given here show that even manufacturers can take advantage of the increased information available from customers to tailor products that at least give the appearance of being customized even if they are simply variations on a common base.

Community

One of the major uses of the Web for both online and offline businesses is to build a network of customers for exchanging product-related information and to create relationships between the customers and the company or brand. These networks and relationships are called communities. The goal is to take a prospective relationship with a product and turn it into something more personal. In this way, the manager can build an environment that makes it more difficult for the customer to leave the "family" of other people who also purchase from the company.

For example, the software company Adobe builds community by devoting a section of its Web site to users and developers. They exchange tips and other information, which binds them more to the company and its brands. By giving the customers the impression that they own this section of the site and by being open to the community about product information, Adobe creates a more personal relationship with its customers. Even companies that sell low-involvement, frequently purchased products attempt to enhance their brands using Web-based communities. For example, Procter & Gamble's Crest brand has a Web site (http://www.crest.com) that invites families to participate in games and obtain information.

Taken together, these relationship programs help to implement the concept of personalization introduced in the introduction. Web-based customer service, loyalty programs, customized products and services, and communities all bring a sense of an individual relationship between company and customer that has been difficult and costly to implement before the commercialization of the Internet.

PRIVACY ISSUES

The CRM system depends on a database of customer information and analysis of that data for more effective targeting of marketing communications and relationship-building activities. There is an obvious trade-off between the ability of companies to deliver customized products and services more effectively and the amount of information necessary to enable this delivery. Particularly with the popularity of the Internet, many consumers and advocacy groups are concerned about the amount of personal information that is contained in databases and how it is being used. Thus, the privacy issue extends all the way through the hierarchy of steps outlined in Figure 1.

This is not a new issue. Direct marketers have mined databases for many years, using analyses based on census tract data, motor vehicle records, magazine subscriptions, credit card transactions, and many other sources of information. With the "in your face" nature of unwanted direct e-mails and the increasing amount of information that is being collected surreptitiously as people browse the Web through ubiquitous "cookies," these concerns have received even more prominence. (For an interesting analysis of how much information can be obtained merely from the user's computer connection to the Internet, visit http://www.privacy.net.) The defining moment in Web privacy occurred in 1999 when the Web ad serving company Doubleclick announced that it was acquiring the direct marketing database company Abacus Direct, with intentions to cross-reference Web browsing and buying behavior with real names and addresses. The public outcry was so strong that Doubleclick had to state that it would not combine information from the two companies.

A study by Forrester Research found a continuum of privacy concerns (Stanley, 2000):

- Simple irritation—primarily the result of unwanted e-mails.
- Feelings of violation—"How do they know that about me?"
- Fear of harm—fears that browsing X-rated sites or booking travel that a consumer does not want others to know about can be revealed, that the Internet can allow others

to access financial or health information, and other concerns.

- Nightmarish visions: Concerns of surveillance by the Internal Revenue Service or "Big Brother" and similar thoughts.

As of this writing, there are eight Internet privacy bills being considered by Congress. The current debate about privacy and the debate in Congress centers around how much control Web surfers should have over their own information. Although many argue that it is in customers' best interest to provide as much data as possible to gain maximum benefit of what the Web has to offer, many disagree. The opponents formalize their arguments in the following two options:

"Opt-in"—In this case, Web users must consent to the collection and use of personal data. This gives the customer more control over his or her own information and helps to build industry confidence. From the marketer's perspective, however, this may substantially reduce the amount of information available in databases.

"Opt-out"—This is the Web version of the direct marketing "negative reply" whereby a customer has to forbid explicitly the collection and use of personal data. This gives more information to marketers and therefore potentially improves the products and services available to customers. The customers bear the loss of control.

These issues are only going to become thornier as the proliferation of wireless devices means more information about customers is available over time. As a result, the main goal of corporate users of Web-based CRM programs should be to build trust between the customer and the company. A large variety of mechanisms for doing this have been proposed (see, for example, Sultan, Urban, Shankar, & Bart, 2002).

METRICS

The increased attention paid to CRM means that the traditional metrics used by managers to measure the success of their products and services in the marketplace have to be updated. Financial and market-based indicators such as profitability, market share, and profit margins have been and will continue to be important. In a CRM world, however, increased emphasis is placed on developing measures that are customer-centric and give managers a better idea of how their CRM policies and programs are working.

Some of these CRM-based measures, both Web- and non-Web-based, are customer acquisition costs, conversion rates (from visitors to buyers), retention/churn rates, same customer sales rates, loyalty measures, and customer share or share of requirements (the share of a customer's purchases in a category devoted to a brand; see Lehmann & Winer, 2002). All of these measures require managers to do a better job acquiring and processing internal data to focus on how the company is performing at the customer level.

THE FUTURE OF CRM

With the increased penetration of CRM philosophies in organizations and the concomitant rise in spending on people and products to implement them, it is clear that we will see improvements in how companies work to establish long-term relationships with their customers. There is, however, a big difference between spending money on these people and products and making it all work: implementation of CRM practices is still far short of ideal. Everyone has his or her own stories about poor customer service and e-mails sent to companies without hearing a response. Despite several years of experience, Web-based companies still did not fulfill many Christmas orders in 2001, and customers continue to have difficulties returning unwanted or defective products. In addition, even those companies who have made large investments in CRM report being dissatisfied with their expenditures on CRM-related hardware and software.

We can expect that the technologies and methodologies employed to implement the steps shown in Figure 1 will improve as they usually do. More companies are recognizing the importance of creating databases and getting creative at capturing customer information. Real-time analyses of customer behavior on the Web for better customer selection and targeting is already here (e.g., Net Perceptions), which permits companies to anticipate what customers are likely to buy. Companies will learn how to develop better communities around their brands, giving customers more incentives to identify themselves with those brands and exhibit higher levels of loyalty.

One way that some companies are developing an improved focus on CRM is through the establishment or consideration of splitting the marketing manager job into two parts: one for acquisition and one for retention. The kinds of skills that are needed for the two tasks are not the same. People skilled in acquisition have experience in the usual tactical aspects of marketing such as advertising and sales. The skills for retention are different because the job requires a better understanding of the underpinnings of satisfaction and loyalty for the particular product category. In addition, time being a critically scarce resource makes it difficult to do an excellent job on both acquisition and retention. As a result, some companies have appointed a chief customer officer (CCO), whose job focuses only on customer interactions.

A possible marketing organizational structure is shown in Figure 6. In this organization, the person overseeing the company's marketing activities, the vice president (VP) of marketing, has both product management and the CCO as direct reports. The CCO's job is to provide intelligence to the VP, from marketing research and from the customer database, for use by product managers in formulating marketing plans and making decisions. In addition, the CCO manages the customer service operation. Although it would perhaps seem more logical for the CCO to report to product management, the reporting arrangement to the VP of marketing is a signal to the company of the prominence of the position. The CCO also interacts with other company managers whose operations may have a direct impact on customer satisfaction.

The CCO at EqualFooting.com, a company offering streamlined purchasing, financing, and shipping services for small manufacturing and construction businesses, has the job of integrating marketing and operations to ensure that customers are satisfied (Manring, 2001). An

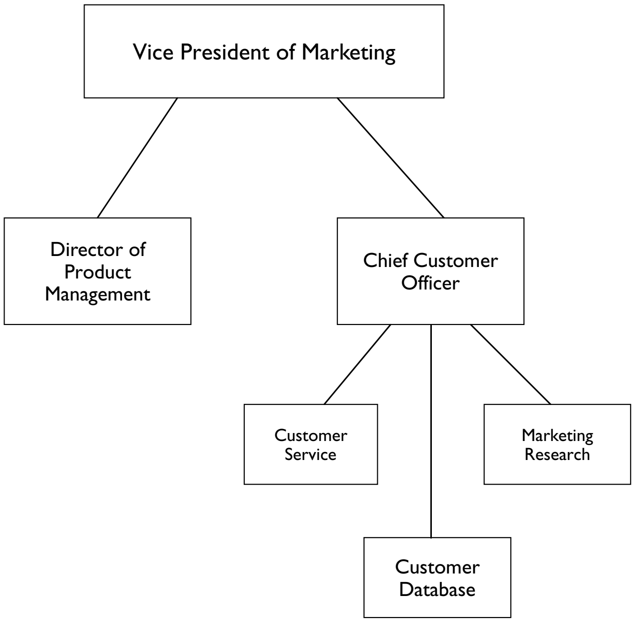


Figure 6: A future marketing organization.

alternative conceptualization is to create two jobs, a customer manager and a capability manager (Pine, Peppers, & Rogers, 2000). The former oversee the relationship with customers while the latter make sure that their requirements are fulfilled.

The notion of customer satisfaction is being expanded to change CRM to CEM—customer experience management. The idea behind this is that with the number of customer contact points increasing all the time, it is more critical than ever to measure the customer's reactions to these contacts and develop immediate responses to negative experiences. These responses could include timely apologies and special offers to compensate for unsatisfactory service. The idea is to expand the notion of a relationship from one that is transaction-based to one that is experiential and continuous.

As with any decision with substantial resource implications, a cost–benefit analysis of CRM investments must be performed. Marketing managers for frequently purchased products such as toothpaste are not as likely to find CRM investments paying out to the extent they will for computer servers, given the differences in difficulties of reaching customers and the profit margins of the respective products. As the Crest community-building effort described earlier shows, however, even toothpaste companies are using the Web to attempt to differentiate their brands from the myriad others appearing in supermarkets and discount stores. This is some evidence that many companies in diverse industries can benefit from the CRM structure.

GLOSSARY

Chief customer officer A manager who is charged with managing a product or company's relationships with its customers.

Churn rate The percentage of a customer base that switches brands in a period.

Clickstream data The data that customers generate from the clicks of the mouse when visiting Web sites.

Customer communities Networks of customers that interact with a company and each other.

A customer database A file of information containing transactions, descriptive information, company contacts, responses to marketing programs, and monetary value to the firm.

Customer experience management (CEM) The management of the experience that a customer has at all of the touch points.

Customer pyramid A hierarchy of groups of customers based on their profitability.

Customer relationship management (CRM) A set of marketing activities and investment in information technology designed to develop long-term relationships with customers.

Direct e-mails E-mails with an advertising message targeting individual customers.

Lifetime customer value (LCV) The present value of the stream of revenues and profits expected to be generated by each customer.

Loyalty and frequency programs Provide rewards to customers for repeat purchasing.

Margin multiple A factor that estimates the LCV when multiplied by the current profit margin that a customer generates.

Mass customization The creation of products and services for individuals or small groups of customers.

Metrics Measures by which the efficacy of marketing programs are judged.

One-to-one Marketing programs that target individual or small groups of customers.

Permission-based e-mails Targeted e-mails in which the customer has given the company permission for contact.

Proactive customer service When a company contacts the customer without a request to do so.

Reactive customer service When the company responds to a customer service request.

Spam Messages from organizations to which a customer has not given permission for e-mail contact.

Touch points All points of contact that a customer has with a firm and its products.

CROSS REFERENCES

See *Consumer Behavior; Data Mining in E-commerce; Data Warehousing and Data Marts; Databases on the Web; Intelligent Agents; Marketing Communication Strategies; Personalization and Customization Technologies; Privacy Law.*

REFERENCES

Anderson, E. W., Fornell, C., & Lehmann, D. R. (1994, July). Customer satisfaction, market share, and profitability. *Journal of Marketing, 58,* 53–66.

Ansari, A., Essegaier, S., & Kohli, R. (2000, August). Internet recommendation systems. *Journal of Marketing Research, 37,* 363–375.

Cigliano, J., Georgiadis, M., Pleasance, D., & Whalley, S. (2000). The price of loyalty. *The McKinsey Quarterly, 4*, 68–77.

Dowling, G. R., & Uncles, M. (1997, Summer). Do customer loyalty programs really work? *Sloan Management Review, 38*, 71–82.

Enrado, P. (2002). Putting the customers first. *Upside,* 52–59.

Glazer, R. (1999, Summer). Winning in smart markets. *Sloan Management Review, 40*, 59–69.

Godin, S. (1999). *Permission marketing.* New York: Simon & Schuster.

Gupta, S., & Lehmann, D. R. (2003). *Customers* as assets. *Journal of Interactive Marketing, 17*, 9–24.

Lehmann, D. R., & Winer, R. S. (2002). *Product management* (3rd ed.). Burr Ridge, IL: McGraw-Hill.

Manring, A. (2001, January). Profiling the chief customer officer. *Customer Relationship Management,* 84–95.

Moe, W. W., & Fader, P. S. (2001, Summer). Uncovering patterns in cybershopping. *California Management Review, 43*, 106–117.

Nail, J. (2000, January). The email marketing dialogue. Technical report. Cambridge, MA: Forrester Research.

O'Connell, V. (2002, July 2). E-mail ads just don't click with consumers. *The Wall Street Journal,* p. B2.

Oliver, R. L. (1997). *Satisfaction: A behavioral perspective on the consumer.* Boston: Irwin McGraw-Hill.

Peppers, D., & Rogers, M. (1993). *The one to one future.* New York: Doubleday.

Pine II, B. J., Peppers, D., & Rogers, M. (2000). Do you want to keep your customers forever? In J. Gilmore & B. J. Pine II (Eds.), *Markets of one.* Cambridge, MA: Harvard Business School Press.

Reichheld, F. F. (1996). *The loyalty effect.* Cambridge, MA: Harvard Business School Press.

Reinartz, W. J., & Kumar, V. (2000, October). On the profitability of long-life customers in a noncontractual setting. *Journal of Marketing, 64*, October, 17–35.

Rust, R. T., Zeithaml, V. A., & Lemon, K. N. (2000). *Driving customer equity.* New York: The Free Press.

Shapiro, C., & Varian, H. R. (1999). *Information rules.* Cambridge, MA: Harvard Business School Press.

Slywotzky, A. J. (2000, January/February). The age of the choiceboard. *Harvard Business Review, 78,* 40–41.

Stanley, J. (2000, May). The Internet's privacy migraine. Technical report. Cambridge, MA: Forrester Research.

Sultan, F., Urban, G., Shankar, V., & Bart, Y. (2002). *Determinants and role of trust in E-business: A large scale empirical study.* Unpublished working paper. Retrieved December 15, 2002, from http://www.venkyshankar.com/home/ViewAbstract.cfm?AbsIndex = 131&ReturnPage = default

Wedel, M., & Kamakura, W. A. (1999). *Market segmentation: Conceptual and methodological foundations* (2nd ed.). New York: Kluwer Academic.

Zeithaml, V. A., & Bitner, M. J. (2000). *Services marketing.* Boston: Irwin McGraw-Hill.

Cybercrime and Cyberfraud

Camille Chin, *West Virginia University*

INTRODUCTION

The phenomenal growth of the Internet continues into the 21st century. With that growth has come tremendous opportunity for education, culture, entertainment, and commerce at the "click of a mouse." However, as the benefits of the Internet increase, the risks have increased as well. We must face the reality that crime has found a home in cyberspace.

Electronic commerce is fast becoming the blueprint for global commerce. For example, 50 million people per day pass through the virtual storefront of eBay; eBay vendors sell goods valued at $12 million every day. The profit-making ability of Internet businesses such as eBay not only boosts the global economy, but also dramatically increases the range of opportunities for criminals to defraud consumers involved in online transactions. The speed and ease of sending e-mail not only facilitate personal and business communication, but also make it easier for computer savvy criminals to use e-mail or keyboard capture devices to obtain the credit card numbers of unwitting consumers.

Cybercriminals are now seizing the opportunity to capitalize on these new business models. While the Internet is a publishing medium that can be used to express our first amendment rights, it is also a medium that can be exploited by forgers and counterfeiters. The challenge that Internet users and lawmakers face is to maximize the commercial and individual benefits of the Internet while minimizing the increasing risk. To minimize risk, we must be educated about cybercrime as well as the resources currently available to protect us from fraudulent online activity. We need to be educated so that we become aware of the gaps that the law does not address with respect to criminal transactions on the Internet. What is cybercrime and cyberfraud? What recourse do we have if cybercriminals victimize us? What can we do to protect ourselves from this new type of crime?

CYBERCRIME

Definition

"Cybercrime" is an evolving concept. Cybercrime constitutes criminal acts, committed by way of computer networks (especially the Internet) that can be accomplished while sitting at a computer keyboard. Such acts include gaining unauthorized access to computer files, disrupting the operation of remote computers with viruses, worms, logic bombs, Trojan horses, and denial of service attacks; distributing and creating child pornography; stealing another's identity; selling contraband; and stalking victims. Cybercrime is a subset of computer crime.

There are many debates today among experts about what constitutes a computer-related crime or cybercrime. Cybercrime generally is understood to include traditional activities such as fraud, theft, or forgery whenever a computer is involved. It can also include a number of new crimes such as cyberstalking (see the online stalking chapter in this encyclopedia).

Cybercrime can also include activities not considered criminal in one jurisdiction, but punishable in another. Since a vast number of criminal acts may constitute cybercrime, this chapter will focus on highlighting the general concept of cybercrime and its classifications, and selected forms of cybersabotage and cyberfraud.

Committing cybercrime is inexpensive if one has the know-how to do it. It is hard to detect if one knows how to erase one's tracks and often is hard to locate in jurisdictional terms, given the geographical indeterminacy of the Internet.

Experts believe that most crimes in the future will involve computers. Recently, the Internet Fraud Complaint Center (IFCC), a public–private partnership among the FBI, the U.S. Department of Justice, and the National White Collar Crime Center, reported that it had received 1,000 complaints of fraud per week since opening in May of 2000. The IFCC anticipates receiving more than 1,000 Internet fraud complaints per day once it becomes fully automated. Cybercrime is growing rapidly across the country, and it is expected that most crimes within the next five years will be related to computers (Internet Fraud Complaint Center, December 2001).

In 1984, Congress enacted U.S. Code (U.S.C.) section 1030 of Title 18, the Computer Fraud and Abuse Act (CFAA), to address computer crimes. The CFAA protects classified information maintained on federal government computers and protects financial records and credit information on government and financial institution computers. Through two amendments in 1986 and 1996, Congress broadened the scope of the CFAA by first extending protection to "federal interest computers" and then by replacing the concept of "federal interest computer" with the phrase "protected computer." A "protected computer" includes any computer "used in interstate or foreign commerce or communication" (Burke, 2001).

Cybercrime Statistics

The most common types of cybercrime, according to statistics compiled by the IFCC, are online auction fraud (48.8%), followed by "nondelivered ordered products" (19.2%), and securities/commodities fraud (16.9%) (IFCC, 2001).

In the IFCC's 2000 Computer Crime and Security Survey, the majority of the respondents were large corporations and government agencies, and 90% of them detected computer security breaches within the last 12 months of the survey. Forty-two percent of the survey's respondents were willing and/or able to quantify their financial loss. The losses from these respondents totaled $265,589,940. Over the past three years, the average annual total was $120,240,180 (Computer Security Institute, 2000).

Of the 1,501 youths, ages 10 to 17, that were surveyed, 20% of them who use the Internet regularly received a sexual solicitation over the Internet in the last year, and 1 in 33 received an aggressive sexual solicitation. Twenty-five percent were unwillingly exposed to pictures of naked people or people having sex; 1 in 17 was threatened or harassed (Finkelhor, 2000).

The most common source of computer crime is a disgruntled employee or former employee. Employees with an average age of 29, holding managerial or professional positions, commit up to 75% of computer crimes (Finkelhor, 2000).

International estimates indicate that cybercrime costs approximately $50 billion annually. More specifically, cybercrime costs the United States more than $5 billion per year. In England, cybercrime is estimated to cost approximately 250 million pounds or $417.7 million annually. In fact, only about 10% of all cybercrimes committed are actually reported and less than 2% result in a conviction. This is primarily due to two reasons. First, businesses and financial institutions believe that they have more to lose by reporting computer security breaches. They argue that customers will lose confidence in the company if business and financial transactions are known to be insecure. Second, a majority of cybercrime victims do not report crimes against them, assuming that law enforcement will provide little or no assistance (Hale, 2002).

TYPES OF CYBERCRIME

Classification of Cybercrime and Cybercriminals

There are two ways to classify cybercrime. One system of classification of cybercrime is the computer as a target or the computer as a tool. The second classification system is to use a profiling approach that distinguishes between insiders and outsiders. The FBI uses the profiling system (FBI, 2000).

Classification 1

Computer as Target. This kind of activity is the wrongful taking of information or the causing of damage to information. A list of specific offenses in this category might include

- Arson (targeting a computer center for damage by fire),
- Extortion (threatening to damage a computer to obtain money),
- Burglary (break-ins to steal computer parts),
- Conspiracy (people agreeing to commit an illegal act on a computer),
- Espionage/sabotage (stealing secrets or destroying competitors' records),
- Forgery (issuing false documents or information via computer),
- Larceny/theft (theft of computer parts),
- Malicious destruction of property (destroying computer hardware or software),
- Murder (tampering with computerized life-sustaining equipment), and/or
- Receiving stolen property (accepting known stolen good or services via a computer).

Computer as Tool. This kind of activity involves modification of a traditional crime by using the Internet in some way. Laws governing fraud apply with equal force regardless if the activity is online or offline, but a few special regulations apply at the federal level:

- Internet fraud (false advertising, credit card fraud, wire fraud, and money laundering);
- Online child pornography, child luring (sexual exploitation, transportation for sexual activity);
- Internet sale of prescription drugs and controlled substances (smuggling, drug control laws);
- Internet sale of firearms (firearms control laws);
- Internet gambling (interstate wagering laws, lottery laws, illegal gambling businesses);
- Internet sale of alcohol (liquor trafficking);
- Online securities fraud (securities act violations);
- Software piracy and intellectual property theft (copyright infringement, trade secrets); and
- Counterfeiting (use of a computer to make duplicates or phonies).

Classification 2

The classification profiles are as follows:

Employees who steal money from the company or harm it some other way. The largest percentage of computer criminals is employees of the particular company being perpetrated. This is usually because they have the easiest access to the computers. Motives vary for employees; they may want to steal some company information and resources for themselves, or they may want to sabotage the company's progress for revenge.

Nonemployees or password crackers ("crackers") who steal or damage its infrastructure in some way. Clients, friends, or suppliers may have access to the company's building, computer system, and/or network. A perfect example of this is a supplier who may have legal access to alter inventory as part of his/her job but also who illegally gains access to suit his/her own needs. Crackers use their skills to gain unauthorized access into computer systems/networks for the fun and challenge of it. They can gain access on-site or can access the systems remotely from their own computers. They may alter files to cause malicious damage for fun and/or for personal profit.

Organized criminals who use a company's information technology for monetary gain (FBI, 2000).

Cybersabotage

Definition

A major subset of cybercrime is cybersabotage, which is the act of using the Internet to undermine the normal functioning of computer systems through the introduction of worms, viruses, or logic bombs. Cybersabotage can be used to promote illegal terrorist activities, to gain economic advantage over a competitor, or to steal programs or data for the purposes of extortion.

Trojan Horses, Viruses, and Worms

Trojan horses, viruses, and worms are all forms of malicious code, which is computer code designed to invade a computer system or to steal information.

Trojan Horses. Trojan horse attacks pose one of the most serious threats to computer security. The name derives from the Greek legend in which the Greeks won the Trojan War by hiding in a huge, hollow wooden horse to get into the fortified city of Troy. Similarly, a computer Trojan horse is a malicious, security-breaking program disguised as something benign such as a screen saver or game.

The most famous Trojan horse was the so-called "Love Bug" in May of 2000. It appeared to be a love letter, but once it was opened, it wreaked havoc by sending itself to everyone in an e-mail address book or Internet relay chat channel, deleting or altering files, and downloading other Trojan horse programs designed to steal passwords.

Many Trojan horses permit password crackers to control a person's computer remotely in order to, for example, use the computer to perform denial of service attacks.

Many people use terms such as Trojan horse, virus, and worm interchangeably, but these terms do not mean the same thing. Trojan horses can be designed to have many functions, such as destroying data, software, and hardware, or transferring a computer virus or a worm. However, a Trojan horse is just as dangerous as a virus or a worm and can easily replicate itself rapidly to harm others. Federal law provides remedies to individuals harmed by Trojan horse programs. Any impairment to the computer system, or data contained therein, is governed by several subsections of section 1030(a) in Title 18 of the United States Code (Colombell, 2002).

Viruses. A malevolent virus is a small piece of programming code usually disguised as benign files that are designed to penetrate and damage personal computers and their contents, operating systems, and computer networks. Programmers write viruses to execute certain functions by altering the original computer program.

Computer viruses are usually designed to spread to other computer users instantly. Viruses transmit themselves as e-mail attachments, as zipped files, or as downloaded files, or they attach themselves to a CD or diskette. Some viruses work instantly; other viruses are initially dormant until the host computer executes the virus' code. Viruses require human interaction to spread, so virus execution occurs when the recipient computer user is tricked into opening an e-mail attachment. Viruses can inflict serious damage, such as erasing hard drives or files, or they are designed to simply taunt the victim with pop-up windows (e.g., pop-up windows saying, "Ha Ha! You are infected!"). There are three main types of viruses:

File infectors: Some file infector viruses attach themselves to program files, .com or .exe files, or they infect programs or .sys, .ovl, .prg, and .mnu files. The virus loads itself when the program loads.

System or boot-record infectors: These viruses infect executable code found in certain areas on a system disk or boot disk. They attach to the DOS boot sector on diskettes or the master boot record on hard disks. A typical scenario is to receive a diskette from an innocent source that contains a boot disk virus. To prevent this type of attack, one should always have a bootable floppy disk.

Macro viruses: Macro viruses infect the Microsoft Word application and typically insert unwanted words or phrases. These are among the most common viruses, and they tend to do the least damage; however, some may argue that they can cause substantial damage (Lo, 2000).

In March of 1999, David Smith, a 28-year-old middle school teacher, authored and launched the Melissa Virus, which harmed over a million computers across the world and caused over $80 million in damages. The virus was hidden in Microsoft Word attachments that appeared to be from a person known to the victim, and once it was activated it e-mailed itself to the first 50 addresses in the victim's Microsoft Outlook e-mail program. Within 48 hours of Melissa's launch, both Intel and Microsoft were forced to shut down their servers.

The federal criminal statute provision that penalizes virus writers such as Smith is 18 U.S.C. section 1030(a)(2). This provision addresses viruses that invade computer systems and sends data from the computer. Another provision is section 1030(a)(5)(A), which prohibits the intentional transmission of viruses disguised as e-mail attachments that damage a computer system. Also, a number of states have adopted statutes criminalizing virus dissemination along with computer fraud and computer intrusions (see Cyber Crime Report at http://www.ifccfbi.gov).

Worms. Worms are special types of viruses that can replicate themselves and use memory but cannot attach themselves to other programs. Unlike viruses, however, worms do not require human interaction and can spread themselves automatically over the network from one computer to the next. Worms are designed to take advantage of automatic file sending and receiving features found on many computers. Malicious worm programs freeze computer monitors and keyboards, disable computers, slow down computers, and use up significant amounts of computer memory.

The Morris Worm, unleashed in November of 1988, was one of the most devastating worms in Internet history. Robert Morris released the worm into a Department of Defense network. It disabled more than 6,000 computers with repair damages per computer ranging from $200 to more than $53,000. In 1989, the Second Circuit Court of Appeals upheld Morris' conviction under the Computer Fraud and Abuse Act, 18 U.S.C. section1030(a)(5)(A) (Burke, 2001).

Computer users can protect themselves from malicious code by using antivirus programs to detect viruses and to prevent their spread. Trojan horses and worms are now more prevalent today than viruses, so antivirus programs are designed to combat rapidly spreading Trojan horses and worms more so than viruses.

The best steps for protection against viruses is to use antivirus software and know the origin of each program or file loaded into the computer or downloaded from e-mail attachments; also check all files periodically and remove any viruses that are found. Be wary of e-mail messages warning of new viruses. Unless the warning is from a recognized source, the warning could be a virus hoax.

Denial of Service Attacks

Denial of service (DoS) attacks, also known as "nukes," "hacking," or "cyberattacks," are incidents in which a user or organization is deprived of the services of a resource they would normally expect to have; in other words, a denial of service attack is an attempt by attackers to prevent legitimate users of a service from using that system. Denial-of-service attacks can essentially disable a computer or network. Usually, the loss of service is the breakdown of network services such as e-mail, or the temporary loss of all network connections and services. In the worst cases, for example, a Web site accessed by millions of people can occasionally be forced to temporarily shut down. Computer system files and programming can also be a casualty of a denial of service attack. DoS attacks are common. The attacks should not be confused with viruses, Trojan horses, worms, and cracking or "hacking."

There are two categories of DoS attacks—operating system attacks and network attacks. Operating system attacks target bugs in specific operating systems and can be fixed with patches. Network attacks exploit networking limitations and may require firewall protection. Denial of service attacks may be part of a larger invasion. Examples of DoS attacks include

attempts to "flood" a network, thereby preventing legitimate network traffic,

attempts to disrupt connections between two machines, thereby preventing access to a service,

attempts to prevent a particular individual from accessing a service, and

attempts to disrupt service to a specific system or person (Lo, 2002).

A denial of service attack can cost the targeted individual or organization considerable time and money. However, these attacks are a type of security breach to a computer system that does not usually result in the theft of information or other security loss. The recent DoS attack on 13 Internet servers around the world illustrates this point. At approximately 4:45 p.m. on Monday, October 21, 2002, these servers were bombarded with 30 to 40 times the normal amount of data. Seven of the 13 servers had periods of zero-response and Internet commerce sites such as eBay, Amazon.com, and Yahoo! also had periods of zero-response. Some have described this incident as the largest denial of service attack in Internet history.

Specific types of denial of service attacks are

a. *Buffer Overflow Attacks:* The most common kind of DoS attack is to overwhelm a network address with more traffic than what the programmers who planned its data buffers anticipated. The attacker may actually be aware of a weakness in the target system and may be trying to exploit it, or the attacker may be testing the system to see if an attack will succeed. Attacks based on the buffer characteristics of programs or systems include these:
   - Sending e-mail messages that have attachments with 256-character file names to Microsoft and Netscape mail programs.
   - Sending a message with a "From" address larger than 256 characters to a user of the certain e-mail programs (Carnegie Mellon Software Engineering Institute/CERT, 1997).

b. *Viruses:* Computer viruses can serve as denial-of-service attacks when there is no specific victim targeted just a host who is unfortunate enough to receive the virus.

c. *Zombies:* A zombie is a computer that has been implanted with a daemon that puts it under the control of a malicious hacker without the knowledge of the computer owner. Malicious hackers use zombies to launch DoS attacks. The hacker sends commands to the zombie through an open port. On command, the zombie computer sends an enormous amount of packets of useless information to a targeted Web site in order to clog the site's routers and keep legitimate users from gaining access to the site. The traffic sent to the Web site is confusing and therefore the computer receiving the data spends time and resources trying to understand the influx of data that has been transmitted by the zombies. Compared to programs such as viruses or worms that can eradicate or steal information, zombies are relatively benign as they temporarily cripple Web sites by flooding them with information and do not compromise the site's data. Such prominent sites as

> Yahoo!, Amazon.com, and CNN.com were shut down in 2000 by zombie DoS attacks. Zombies are also referred to as *zombie ants* (Colombell, 2002).

The Computer Fraud and Abuse Act has several provisions that regulate denial of service attacks. DoS attacks can be a serious federal crime under the National Information Infrastructure Protection Act of 1996, with penalties that include imprisonment. At a minimum, DoS offenders usually lose their Internet service provider accounts or are suspended if they used school resources to launch the attack. However, DoS attacks are not explicitly criminalized in many state codes (FBI, 2000).

Social Engineering

Social engineering is a password cracker's use of psychological tricks on legitimate users of a computer system/network to gain the information (user names and passwords) he/she needs to gain access to a computer system/network. Social engineering is basically a new way to commit fraud in the information age or to "pull a con job," hacker style. The object is to get information or access to systems that are normally only used by privileged users. The following excerpt illustrates examples of social engineering:

> A woman calls a company help desk and says she's forgotten her password. In a panic, she adds that if she misses the deadline on a big advertising project her boss might even fire her. The help desk worker feels sorry for her and quickly resets the password—unwittingly giving a hacker clear entrance into the corporate network.
>
> Meanwhile, a man is in back of the building loading the company's paper recycling bins into the back of a truck. Inside the bins are lists of employee titles and phone numbers, marketing plans and the latest company financials. All free for the taking. (Gaudin, 2002)

Hackers, and possibly even corporate competitors, are breaching companies' network security every day. The latest survey by the Computer Security Institute and the FBI shows that 90% of the 503 companies contacted reported break-ins within the last year.

What may come as a surprise, according to industry analysts and security experts, is that not every hacker is sitting alone with his/her computer hacking his/her way into a corporate VPN or running a program to crack executives' passwords. Sometimes all hackers have to do is call up and ask (Gaudin, 2002).

Social engineering has existed in some form or another since the beginning of time, primarily because most people are helpful and trusting. It is human nature. The Love Bug virus illustrates this; it was designed to exploit the psychological need and/or want of human beings to be loved. Other social engineering tools used by password crackers include the telephone, e-mail viruses, "dumpster diving," and the use of "snail mail"—mailing a fraudulent survey that asks delicate questions and offers a cash reward for completion.

So with the immeasurable security threat that social engineering brings to the computing community, why is very little ever said about it? In general, victims view social engineering as an attack on their intelligence, and apparently victims do not want to be considered "ignorant." The reality is that a large number of computer users are, and will continue to be, susceptible to social engineering attacks.

Although the methods used by social engineers rely on the same principle, the disguises of the password crackers may vary greatly, depending on the cracker's level of skill and the type of information he or she seeks. One common method used is for the attacker to pretend he/she is new to the system and needs assistance with gaining access. The role as a new person ("newbie") is easy for a potential password cracker to assimilate. The cracker can easily pretend not to know much about a system and still retrieve information. This ruse is commonly used when the attacker is unable to research enough about the company or finds enough information to undermine computer security. A simple method of this technique is for the cracker to call a secretary for the company and pretend that he or she is a new temporary agent having trouble gaining access into the system. The secretary (or other legitimate user) may be inclined and proud to be able to offer help to the new person on the job. The user may simply give out the guest account name and password or may even go into detailed instructions on log-in procedures for different departments. Once the intruder is in a guest account, however, he or she may be able to access other (more important) accounts from there.

Another guise used by social engineers is to pose as computer aides or helpers and try to gain information as the potential victim fixes the computer. This technique, however, relies on the assumption that there is something wrong with the computer system. When the social engineer poses as a helper, the legitimate user often is less suspicious and more willing to answer inquisitive questions. Another role for the attacker to assume is that of a system operator for the network itself. The potential password cracker will pretend that an error in all the accounts has been made and that he or she needs to reset the accounts. To do that, the attacker needs the old passwords of the users. If the employee is naive enough, he or she will divulge the information, thinking that they are doing their company a service (Nelson, 2000).

The most famous social engineer to date is Kevin Mitnick. Mitnick served almost five years in prison for breaking into computers, stealing data, and abusing electronic communication systems. Mitnick said in a speech, "You try to make an emotional connection with the person on the other side to create a sense of trust. That is the whole idea: to create a sense of trust and then exploit it. Through social engineering, I gained the ability to obtain any number, listed or unlisted. This really came easy to me—manipulating the Telephone Company" (Lemos, 2000).

What can be done to protect a person or business from social engineering? First, potential victims need to be aware that it exists. Individuals must be trained to avoid giving out any more information than necessary if anyone solicits sensitive information. In other words, if anyone asks for a password, one must avoid divulging one's password information or divulging anyone else's password to anyone for any reason.

If a person or business suspects a social engineering attack, or is the victim of a social engineering attack, the victim should immediately notify the person(s), or department(s) responsible for computer/network security. The only way that these criminals can be apprehended is if more victims quickly and accurately report these crimes as they occur.

Cyberfraud

Definition

Cyberfraud is any type of fraud that occurs on the Internet. The most common forms are online auction fraud, including nondelivery of paid products purchased through online auctions, identity theft, and credit card theft. Other common forms of cyberfraud include nondelivery of merchandise or software bought online from non-auction sites, data break-ins, and online securities fraud.

Internet Auction Fraud

What is an Internet auction? An Internet auction site is a virtual marketplace that requires sellers to register and create a user identification tag. The seller then lists the items they wish to sell on the Web site, usually including a digital picture of the item and a written description. Bidders are required to register and create user identification tags before they place bids on items they wish to purchase. Potential buyers bid against each other until the auction ends. The seller determines the time and date for the end of the auction. When the auction expires, the highest bidder wins (Federal Trade Commission, 2000b). The winner sends payment, including shipping and handling charges, to the seller, either through credit card payments, money orders, checks, or online payment methods such as Western Union BidPay, PayPal, and eBay's Billpoint. Once the payment is received, the seller ships the auctioned item(s) to the buyer to complete the transaction.

What is Internet auction fraud? There are seven types of Internet auction fraud, and most of them involve fraud on the part of the seller: nondelivery of goods, triangulation, misrepresentation, shill bidding, selling black market goods, fee stacking, and bid shielding or multiple bidding (IFCC, 2001).

The first category is nondelivery of goods. This occurs when the seller knowingly auctions off an item that does not exist. The seller places a description and digital picture of the nonexistent item on an auction Web page, receives payment from the highest bidder at the end of the auction, and never delivers anything to the buyer. In addition to duping the buyer out of the winning bid price, the seller also gets access to the buyer's credit card number and name, which can be used for identity theft crimes.

Triangulation schemes also combine identity theft with Internet auction fraud, and victimize online merchants and unwitting bidders. The perpetrator steals the credit card number and identity of another and uses that information to purchase goods from online merchants. The thief then offers the purchased goods for sale on an online auction site and sells the goods to the highest bidders.

Misrepresentation occurs when the seller places false or misleading information about the item offered for sale on the auction site to make the item appear to be more valuable than it actually is.

Shill bidding is when the seller or an accomplice of the seller uses another user identity to bid against potential buyers. The result is that the seller or the seller's accomplice purposely drives up the price of the item being auctioned, to the detriment of legitimate bidders.

Perpetrators who auction black market goods for sale usually place illegally obtained goods, such as pirated CDs, software, and videos, onto Internet auction Web sites without the appropriate packaging, instructions, or warranty information.

Another type of illegal auction fraud identified by the IFCC is fee stacking. Fee stacking occurs when the seller drives up the total price of the merchandise auctioned by adding separate fees for shipping containers, for handling, and for shipping instead of listing a flat rate for shipping and handling.

Bid shielding or multiple bidding is similar to shill bidding except that it involves fraud on the part of the buyer. The buyer creates several user identities and uses them to place multiple bids for the same item. By driving up the price with multiple aliases, the buyer drives other potential buyers away, and then the buyer retracts the high bids placed just before the end of the auction. The result is that the buyer wins the auction at a lower bid price.

What are examples of Internet auction fraud? In May of 2001, an eBay seller from Arlington, Virginia, was arrested for defrauding 20 victims out of approximately $6,000. The seller sold 20 first editions of the Harry Potter book series containing fraudulent autographs. The perpetrator placed digital pictures of the books along with a description in which he claimed that the books were worth $3,000 because they were legitimate first editions or first prints autographed by the author of the Harry Potter series (Duffy, 2001).

In February of 2002, three men of Seattle, Washington, were arrested and charged with wire fraud for running a shill-bidding ring on eBay. The seller and two coconspirators were accused of manufacturing over $1.3 million in high bids on highly collectable Rene Lalique glass (David Steiner, 2002).

In May 2002, an online auction seller was indicted for defrauding over 30 victims out of more than $25,000. Over a five-year period, the perpetrator offered computer laptops for sale and instructed the auction winners to send payment to a mailbox address. The seller never delivered goods to the victims (Ina Steiner, 2002).

What can be done to protect oneself from Internet auction fraud? Buyers should do the following: Get as much information as possible about the seller before bidding on any items the seller offers at auction. Buyers should also try to get the seller's name, physical address, and phone number. Potential victims should be particularly cautious if the seller only lists a post office box address instead of a physical address, if the seller is located outside of the United States, or if the seller only lists an e-mail address. Before engaging in these types of transactions, buyers should check the seller's feedback ratings posted on the auction site from prior buyers, but be wary of shill feedback posted by the seller or the seller's friends. Before bidding, buyers should determine what form of payment the seller will accept and be especially cautious if the seller accepts only money orders or cashier's checks and does not accept credit cards. It is safer to use credit cards, escrow services, or online payment services such as PayPal or use cash on delivery C.O.D. through the U.S. Postal Service (*Bidder beware,* 2002).

Buyers should also be sure to get information about the shipping and handling charges to be added to the price of the item before bidding. The more accurate information a potential buyer has prior to bidding, the more likely one is to protect oneself from this type of fraud; buyers should take time to learn exactly how the online auction works before bidding. Finally, buyers should never give out personal identifying information such as social security numbers or driver's license numbers to sellers.

What efforts have been made to protect buyers from Internet auction fraud? Presently, there is no regulation of Internet auction sites other than self-regulation.

Online auction sites usually place disclaimers throughout the site that are structured to minimize legal liability for any fraudulent activity occurring on the site. Internet auction sites are more similar to flea markets than to traditional auctions because these sites simply provide a venue for buyers and sellers to buy and sell their wares; unlike many traditional auctions, internet auction sites never possess title to any of the auctioned items.

Other than using the host site's feedback rating system, the only recourse that a victimized consumer has is to file a complaint with agencies such as the Federal Trade Commission, the Internet Fraud Initiative established by the Department of Justice, the IFCC, or the Better Business Bureau Online (Albert, 2002).

Internet Identity/Credit Card Theft

What is identity theft? Identity theft is the act of stealing the identifying information of others to access their financial resources through fraud and impersonation.

How does someone steal one's identity? First, the identity thief or impostor steals personal information such as one's name, birth date, social security number, driver's license number, credit card numbers, bank account numbers, mother's maiden name, and any other identifying information. Second, the identity thief uses the stolen personal information to commit fraudulent acts by assuming the victim's identity and withdrawing money from the victim's bank accounts, or submitting false applications for credit cards, auto loans, or bank loans. Sometimes the impostor uses the victim's personal information to open bank accounts, set up utility or cellular phone accounts, purchase homes, lease apartments and cars, or obtain driver's licenses and passports that contain the victim's name and personal information but with the perpetrator's picture in place of the victim's (Federal Trade Commission, 2002a).

How does the identity thief obtain a victim's personal information? Thieves watch victims as they enter credit card numbers, calling card numbers, and PINs ("shoulder surfing"). They listen as victims give their credit card information over the telephone. They steal a victim's mail and go through their garbage for bank or credit card statements or medical records. They break into a victim's home or automobile to steal personal information, and they "break in" or use password cracking skills or keystroke mirroring software to gain access to a victim's personal computer. Thieves use electronic devices to download the credit or debit card information of their victims ("skimming") (Hoar, 2001), or they use Internet databases to select their victims (Sabol, 1999).

What is the impact of identity theft on the victim? It may take months or even years before a victim realizes that their identity has been stolen. During that time, the perpetrator borrows and spends thousands of dollars using the victim's name and creditworthiness. Usually the defrauded creditor initiates collection action or even lawsuits against the victim, and the victim's credit rating and reputation is destroyed. Sometimes the victim is arrested for crimes committed by the impersonator. In some cases, victims have lost job opportunities or have even been fired.

What are examples of identity theft: In September 1997, the United States Secret Service brought an identity theft case against a couple from Maryland. They pled guilty to incurring $100,000 of debt under the stolen identities of victims that they selected from Internet databases (Saunders & Zucker, 1999).

In July and October of 2000, two members of an identity theft ring were prosecuted in the federal district court in Oregon. They hacked into personal computers and Web sites with keystroke mirroring software designed to capture credit card information, and they burglarized automobiles, stole mail, and searched the garbage of their victims to obtain identity information. The thieves used the stolen information to make approximately $400,000 in purchases on 400 fraudulently obtained credit card accounts (Hoar, 2001).

In August of 2000, two defendants were convicted in the federal district court in Delaware of obtaining the names and social security numbers of high-ranking military officers from an Internet Web site. They then used the information to incur approximately $300,000 in debt from fraudulently obtained bank and corporate lines of credit as well as credit cards (Hoar, 2001).

Last year, over 250,000 customers of an online bank were duped into providing bank account information to criminals. The thieves copied the bank's Web site and then manufactured a fraudulent e-mail message, purportedly sent from the bank, requesting that customers go to the link (to the copied Web site) and reregister account information that had been lost due to "technical problems." (Cronin & Weikers, 2002).

What laws exist to protect victims of identity theft? Congress enacted the Identity Theft and Assumption Deterrence Act on October 30, 1998. The Act mandates that the theft and misuse of personal information with the intent to commit an unlawful act be subject to a maximum fine of $250,000 and up to 25 years' imprisonment. As the agency primarily responsible for enforcing the Act, the Federal Trade Commission records and tracks complaints of identity theft, coordinates investigation of complaints with law enforcement agencies, and educates the public about identity theft (Saunders & Zucker, 1999). Several state legislatures have also enacted statutes designed to criminalize identity theft.

What can be done to protect oneself from identity theft? The victim(s) should contact the defrauded creditors, major credit bureaus, the social security administration, local police, and postal inspectors. The victim(s) should also follow up with a written correspondence via certified mail, return receipt requested. Victims may also want to contact the National Crime Information Center (Sabol, 1999).

Internet Investment Fraud

What constitutes Internet investment fraud? The most prevalent form of fraudulent activity related to investment activity on the Internet is market manipulation. In Ernst & Ernst v. Hochfelder, 425 U.S. 185, 195 (1976), the U.S. Supreme Court defined market manipulation as activity that "connotes intentional or willful conduct designed to deceive or defraud investors by controlling or artificially affecting the price of securities."

There are several types of market manipulation that occur over the Internet. These include, the posting of misleading information on message boards or "Web hoax" cases; the "accidental" delivery of private insider information via e-mail to "unintended recipients," and the posting of misleading investment information in chatrooms through "pump and dump" schemes or "stock guru" cases and/or "momentum trading" schemes, or "free stock offer" cases (Walker & Levine, 2001).

What are examples of Internet investment fraud? On August 25, 2000, misleading information about Emulex Corporation was posted in the form of a press release to an online news service. The fake press release indicated that the CEO and President of Emulex resigned, that the Securities Exchange Commission (SEC) was investigating accounting irregularities at the company, and that the company had sustained fourth quarter losses. For about an hour and a half, two different investment news services ran headlines based on the false press release. Even after the Web hoax was reported, Emulex stock was trading down 6 3/8 points.

On September 20, 2000, the SEC charged a 15-year old boy from New Jersey with civil fraud for making $272,826 of illegal profits based on a "stock guru" scheme. He created several aliases and brokerage accounts to purchase a large block of stock. He then sent misleading e-mail messages pumping up the value of the stock to several investment message boards. Within 24 hours, the perpetrator sold the large block of stock and gained profits between $11,000 and $74,000 on each trade (Walker & Levine, 2001).

On September 6, 2000, the SEC charged the creator of an e-mail list referred to as the "Unity List" with violations of section 17(a) of the Securities Act of 1933, and sections 10(b) and 10(b)(5) of the Securities Exchange Act of 1934. The perpetrator sent thousands of e-mails to the names on the e-mail list indicating that he intended to purchase large blocks of stock from certain companies and urging the e-mail recipients to buy stock in those same companies. The perpetrator owned shares in those companies beforehand and sold those shares as soon as the momentum he created in that stock caused the price per share to increase. He made over $12,000 in profits from this momentum trading scheme (Bell, 2002).

The SEC regulates the securities market and enforces the provisions of both the Securities Act of 1933 and the Securities Exchange Act of 1934. The Securities Exchange Act of 1934 applies to online market manipulation and regulates secondary trading. The 1934 Act contains antifraud provisions that make deceptive and manipulative conduct in connection with the purchase or sale of securities unlawful. Civil and criminal law suits may be initiated under the antifraud provisions. Each state also has securities enforcement agencies and state statutes designed to regulate securities fraud (Hittle, 2001). The SEC's Division of Enforcement (DOE) initiates investigations of securities law violations and files civil lawsuits against violators of the securities laws; the DOE also issues injunctions, and in some cases, initiates criminal prosecution. In addition, private individuals may initiate civil law suits.

In July 1998, the SEC established the Office of Internet Enforcement to investigate Internet investment fraud, and to oversee the online Enforcement Complaint Center and the Cyberforce, a team of investigators, lawyers, and accountants who surf the Internet looking for evidence of online securities fraud.

What steps can potential investors take to protect them from Internet investment fraud? Investors can educate themselves to recognize the warning signs of Internet investment fraud by taking advantage of the resources offered by educational agencies such as the SEC's Office of Investor Education and Assistance.

Potential investors can also access large amounts of information about companies by using the Internet. They can look for any evidence of insider trading and any evidence of inconsistencies between past company statements and current statements. Message boards can provide insight and additional information about companies. If an investor believes he or she has been defrauded then he or she may use the Internet to get information on any recently filed securities fraud cases against the company.

CONCLUSION

The advancement of Internet commerce has enabled many individuals around the world to become entrepreneurs. However, criminally minded individuals have used the Internet as a means to commit crime. Different variations of cybercrime and cyberfraud, Internet auction fraud and identity theft, for example, have increased dramatically in the past five years and will continue to increase.

Cybercriminals now have an arsenal of sophisticated tools to further their goals. Criminals can engage in corporate sabotage by planting Trojan horses, worms, and viruses in computer networks. Password crackers can use long established methods of fraud to extract delicate information from unwitting victims through social engineering. Cybercriminals can completely shut down businesses and potentially damage the Internet with denial of service attacks. In fact, the most destructive denial of service attack in Internet history occurred at the end of the business day on Monday, October 21, 2002. Thirteen computer servers around the world were attacked with 30 to 40 times the normal amount of data. Sites such as eBay, Amazon.com, and Yahoo! were temporarily inaccessible as a result of this attack. The incident highlighted the need for government and industry to stay ahead of those individuals who attempt to sabotage Internet security and Internet commerce.

The incidence of cyberfraud has also increased dramatically over the past five years. Online auctions present a myriad of opportunities for cybercriminals to engage in fraudulent transactions. The ease with which anyone can use the Internet to obtain personal information about consumers explains the sharp increase in identity theft and credit card fraud. White-collar criminals can now use the Internet, e-mail, chat rooms, and bulletin boards to manipulate markets on a much larger scale than ever before.

All Internet users should take advantage of the wealth of opportunities that the Internet affords us; however, Internet users should also take steps to prevent and stop the abuse and exploitation by cybercriminals. The only way to achieve this goal is to promote cybercrime prevention through wide-scale education and training of the participants of e-commerce.

GLOSSARY

Antivirus software A class of program that searches the computer hard drive and floppy disks for known or potential viruses.

Back door A vulnerability intentionally left in the security of a computer system or its software by the software programmers.

Black hat A term used to describe a "hacker" who has the intention of causing damage or stealing information.

Buffer A buffer is a data area shared by hardware devices or program processes that operate at different speeds or with different sets of priorities. The buffer allows each device or process to operate without being delayed by the other. For a buffer to be effective, the size of the buffer and the algorithms for moving data into and out of the buffer need to be considered by the buffer designer. Like a "cache," a buffer is a "midpoint holding place" but exists not so much to accelerate the speed of an activity as to support the coordination of separate activities. This term is used both in programming and in hardware. In programming, buffering sometimes implies the need to screen data from its final intended place so that it can be edited or otherwise processed before being moved to a regular file or database.

Cache (pronounced "cash") **1.** A storage area on the computer motherboard that keeps frequently accessed data or program instructions readily available. **2.** In a browser, a section of the hard drive that is set aside for storing recently accessed Web pages.

Cracker Password crackers—"crackers"—are often mistakenly called "hackers." Crackers are the so-called "bad guys" who seek to "crack" or gain unauthorized access to computers, typically to do malicious acts, e.g., to steal credit card information or destroy computer files. Crackers might accomplish this by writing a program called a "virus," "worm," or "Trojan horse." Alternatively, they may exploit weaknesses in the computer's operating system in order to gain entry. Many crackers will install a "backdoor" that allows the cracker to "remote control" the user's computer over the Internet, for example, to distribute child pornography or to perform a "denial of service attack" against somebody else.

Cracking The process used by "crackers" (see above) attempting to overcome a security measure.

Daemon (pronounced "demon" or "damon") A program, usually on a computer running the UNIX operating system, that serves some obscure function (such as routing e-mail to its recipients) and usually has a very limited user interface.

Denial of service attack An attack that causes the targeted computer system to be unable to fulfill its intended function.

Firewall A system designed to prevent unauthorized access to or from a private computer network. Firewalls can be implemented through both hardware and software, or a combination of both. Firewalls are frequently used to prevent unauthorized Internet users from accessing private networks connected to the Internet, especially intranets. All messages entering or leaving the intranet pass through the firewall, which examines each message and blocks or denies entry to those that do not meet the specified security criteria. A device designed to enforce the boundary between two or more networks, limiting access.

Hacker When used properly, this term refers to an elite breed of so-called "good guys," who are talented computer programmers. They enjoy solving challenging problems or exploring the capabilities of computers. True hackers subscribe to a code of ethics and decry the illegal and immoral activity of crackers (defined above). When the press uses "hackers" to describe authors of virus programs or criminals who use the Internet to commit theft or vandalism, it is not only inaccurate, but also interpreted as insulting to those who define themselves as true hackers.

Hacking The original term referred to the learning of programming languages and computer systems; now associated with the process of bypassing the security systems on a computer system or network.

Internet service provider A company that provides access to the Internet for individuals or other companies.

Internet relay chat (IRC) A "chat" system developed by Jarkko Oikarinen in Finland in the late 1980s. IRC has become very popular as more users connect to the Internet because it enables those connected at any location on the Internet to join in live typed discussions. Unlike older chat systems, IRC is not limited to two participants.

Logic bomb A delayed-action computer virus or Trojan horse. A logic bomb, when "exploded," may be designed to display or print a spurious message, delete or corrupt data, or have other undesirable effects.

Network In information technology, a network is a series of points or nodes interconnected by electronic communication paths. Networks can interconnect with other networks and contain subnetworks. The "Internet" is a network of networks, i.e., an internetwork.

Node In a computer network, a node is a connection point, usually a computer.

Password sniffing The process used by a program to examine data traffic for the purpose of finding passwords to use later in masquerading attacks.

Trojan horse An apparently innocuous program that contains code designed to surreptitiously access information or computer systems without the user's knowledge. For example, a harmless holiday greeting card delivered via e-mail, when opened, executes a program to copy the user's files and transmit them to the attacker through the Internet.

Virus A computer program designed to replicate or copy itself and spread the copies itself from one machine to another without the aid, and often without the knowledge, of the user.

Virus hoax A virus hoax is a false warning about a computer virus.

Worm A worm is a self-replicating virus that does not alter files but resides in active memory and duplicates itself. Worms use parts of a computer operating system that are automatic and usually invisible to the user. It is common for worms to be noticed only when their uncontrolled replication consumes system resources, slowing or halting other tasks.

CROSS REFERENCES

See *Computer Viruses and Worms; Cyberlaw: The Major Areas, Development, and Provisions; Cyberstalking; Cyberterrorism; Denial of Service Attacks; Digital Identity; Firewalls; International Cyberlaw; Legal, Social and Ethical Issues; Privacy Law.*

REFERENCES

Albert, M. (2002, Summer). E-buyer beware: Why online auction fraud should be regulated. *American Business Law Journal, 39,* 575.

Bell, B. (2002, June). The evolving use of the internet in connection with securities litigation. *Cyberspace Lawyer,* 7(4), 2.

Bidder beware: Towards a fraud free marketplace—Best practices for the online auction industry (2002, May 31). Retrieved October 24, 2002, from http://www.law.washington.edu/lct/files/auction%20best%20practices.doc

Burke, E. (2001, January). The expanding importance of the Computer Fraud and Abuse Act. Retrieved October 24, 2002, from http://www.gigalaw.com/articles/2001-all/burke-2001-01-all

Carnegie Mellon Software Engineering Institute/CERT (1997, October). Denial of service attacks. Retrieved October 24, 2002, from http://www.cert.org/tech_tips/denial_of_service.html

Computer Security Institute (2000). CSI/FBI 2000 computer crime and security survey. Retrieved October 24, 2002, from http://www.pbs.org/wgbh/pages/frontline/shows/hackers/risks/csi-fbi2000.pdf

Colombell, M., & Rich, J. L. (2002, Spring). The legislative response to the evolution of computer viruses. *Richmond Journal of Law and Technology, 8,* 1–59.

Cronin, K., & Weikers, R. (2002). *Data security and privacy law: Combating cyberthreats* (section 1:7—Identity theft). Eagan, MN: WestGroup.

Duffy, M. (2001, April 9). *Alleged Harry Potter fraud on eBay.* Retrieved October 24, 2002, from http://www.auctionwatch.com/email/print.html?ret = /awdaily/dailynews/april01/1—040901

Ernst & Ernst v. Hochfelder. (1976). 425 U.S. 185, 195.

Federal Bureau of Investigation (FBI) (2000, February). *Statement for the record of Louis J. Freeh, Director, Federal Bureau of Investigation, on cybercrime before the Senate Committee on Appropriations Subcommittee for the Departments of Commerce, Justice, State, the Judiciary, and Related Agencies.* Retrieved October 24, 2002, from http://www.fbi.gov/congress/congress00/cyber021600.htm

Federal Trade Commission (2002a). *ID theft: When bad things happen to your good name.* Retrieved October 24, 2002, from http://www.ftc.gov/bcp/conline/pubs/credit/idtheft.htm

Federal Trade Commission (2000b). *Internet auctions: A guide for buyers and sellers.* Retrieved October 24, 2002, from http://www.ftc.gov/bcp/conline/pubs/online/auctions.htm

Finkelhor, D. (2000). *Online victimization: A report on the nation's youth.* Retrieved October 24, 2002, from http://www.unh.edu/ccrc/pdf/Victimization_Online_Survey.pdf

Gaudin, S. (2002, May 10). *Social engineering: The human side of hacking.* Retrieved October 24, 2002, from http://itmanagement.earthweb.com/secu/article.php/1040881

Hale, C. (2002, September). Cyber crime: Facts & figures concerning the global dilemma. *Crime & Justice International, 18*(65), 5–6, 24–26.

Hittle, B. (2001, December). An uphill battle: The difficulty of determining & detecting perpetrators of Internet stock fraud. *Federal Communication Law Journal, 54,* 165.

Hoar, S. (2001, Winter). Identity theft: The crime of the new millenium. *Oregon Law Review, 80,* 1423.

Internet Fraud Complaint Center (2001, December). *IFCC annual Internet fraud report.* Retrieved October 24, 2002, from http://www1.ifccfbi.gov/strategy/IFCC_2001_AnnualReport.pdf

Lemos, R. (2000, July 16). Mitnick teaches "social engineering." *.ZDNet News.* Retrieved October 24, 2002, from http://zdnet.com.com/2100-11-522261.html?legacy = zdnn

Lo, J. (2000, May 6). *Trojan horse or virus?* Retrieved October 24, 2002, from http://www.irchelp.org/irchelp/security/trojanterms.html

Lo, J. (2002, January 6). *Denial of service or "nuke" attacks.* Retrieved October 24, 2002, from http://www.irchelp.org/irchelp/nuke/

Nelson, R. (2002, April). *Methods of hacking: Social engineering.* Retrieved January 11, 2003, from The Institute for Systems Research, University of Maryland Web site at http://www.isr.umd.edu/gemstone/infosec/ver2/papers/socialeng.html

Sabol, M. (1999). The Identity Theft & Assumption Deterrence Act of 1998: Do individual victims finally get their day in court? *Loyola Consumer Law Review, 11,* 165–171.

Saunders, K., & Zucker, B. (1999, Spring). Counteracting identity fraud in the information age: The Identity Theft & Assumption Deterrence Act. *Journal of Law and Public Policy, 8,* 661–675.

Steiner, D. (2002, February 8). *Authorities crack shill bidding ring on eBay. Auctionbytes News Flash, 327.* Retrieved October 24, 2002, from http://www.auctionbytes.com/pages/abn/y02/m02/i08/s02

Steiner, I. (2002, May 31). *Online auction seller indicted on mail fraud charges.* Retrieved October 24, 2002, from http://www.auctionbytes.com/pages/abn/y02/m05/i31/s01

Walker, R., & Levine, D. (2001, Summer). "You've got jail": Current trends in civil & criminal enforcement of Internet securities fraud. *American Criminal Law Review, 38,* 405.

FURTHER READING

Web Sites

Black Hat: http://www.blackhat.com/presentation/bh-usa-02/bh-us-02-shinder-cybercrime.ppt

The Berkman Center for Internet and Society at Harvard Law School: http://eon.law.harvard.edu/ilaw/Cybercrime

Computer Crime and; Intellectual Property Section (CCIPS) of the Criminal Division of the U.S. Department of Justice: http://www.cybercrime.gov/

The Criminal Justice Resources, Michigan State University Libraries: http://www.lib.msu.edu/harris23/crimjust/cybercri.htm

Davis Logic: http://www.davislogic.com/cybercrime.htm

Internet Tips and Secrets: http://Internet-tips.net/Security/social.htm

The Jargon Dictionary: http://info.astrian.net/jargon/

National Center for Victims of Crimes: http://www.ncvc.org/stats/cyber.htm

Netlingo: The Internet Dictionary: http://www.netlingo.com

North Carolina Wesleyan: http://faculty.ncwc.edu/toconnor/315/315lec12.htm

TechTarget: http://searchsecurity.techtarget.com/

TechTV/Cybercrime: http://www.techtv.com/cybercrime/

The Third Branch, Newsletter of the Federal Courts: http://www.uscourts.gov/ttb/sept00ttb/sept00.html

Articles

Brooks, R. (1998, Spring). Deterring the spread of viruses online: Can tort law tighten the 'net'? *Litigation Review, 17,* 343–391.

Byers, S. (2001, Fall). The Internet: Privacy lost, identities stolen. *Brandeis Law Journal, 40,* 141–162.

Internet Fraud Complaint Center (2001, May). *Internet auction fraud.* Retrieved October 24, 2002, from http://www1.ifccfbi.gov/strategy/AuctionFraudReport.pdf

Sinrod, E., & Reilly, W. (2000, May). Cyber-Crimes: A practical approach to the application of federal computer crime laws. *Santa Clara Computer & High Technology Law Journal, 16,* 177.

Cyberlaw: The Major Areas, Development, and Provisions

Dennis M. Powers, *Southern Oregon University*

INTRODUCTION

As the number of Internet users, Web connections, and personal computers increased exponentially, controversies and legal problems also accelerated in cyberspace without any specific statutes or case law to govern conflicts. Despite solid preexisting legal foundation, no information medium ever had such an enormous appetite to leapfrog geographic territories and laws, in turn creating intense pressures on that system.

If somebody "ripped off" another's slogan or logo in "pre-Net" California, it was possible that a business located in Chicago would never know the difference. If one person wrote a defamatory article in a local Florida newspaper about someone in Oregon, the defamed person at that time could have died before reading that particular printed statement. Once the Internet came into existence, however, anyone with an Internet connection—whether living in Florida or in France—could stumble across that slogan or posting. Large organizations were caught napping when more nimble entrepreneurs registered their trademarks as domain names, and then offered to sell those registrations back for outrageous sums of money. And who could ever have predicted the rise of mass-copying technology such as Napster that bypassed the copyrights of the musicians, composers, and recording studios?

Then add in the wide differences among the laws of wide numbers of states and other countries. Whether entered into by e-mail or not, a contract under certain facts could be fine in Georgia but void in California, and a copyright claim upheld in Japan, but not in the United States with its differing laws and "fair use" exception. Over time, court decisions confirmed existing legal concepts as applicable to cyberspace, and specific statutes were enacted to fill in the gaps. These basic legal concepts with later refinements proved adaptable to the Internet technology of the new millennium, just as they had during the dawning of new technologies in the past century and as unfolds during the course of this chapter.

COPYRIGHT LAW

Your computer is a worldwide copying machine, and the Internet made it extraordinarily easy for nearly anything to be instantly copied, e-mailed, and printed out anywhere, regardless of the true copyright holder's rights. In response, the United States took the lead to enact legislation that complemented the basic law of copyrights and met this tension between strong competing interests.

The first major step taken was its enactment of the U.S. Digital Millennium Copyright Act of 1998 (DMCA). The online service provider section of the DMCA establishes the procedures for copyright owners to contact service providers with their complaints over a subscriber's improper online use of copyrighted material. This act mandates providers to remove materials used improperly once they reasonably determine there are, in fact, copyright infringements as alleged. If the subscriber files a counterprotest, however, then the provider must repost the material unless the complainant files a lawsuit against the infringer for copyright infringement over the offending use.

In effect, this legislation grants copyright owners an administrative tool to remove infringing material without having to litigate the problem, as well as a "safe harbor" against liability for U.S. online service providers. The definition of "service providers" is broad and includes not only Internet service providers (ISPs), Web hosting companies, wire and fiber transmission entities, and router services, but also corporations, universities, municipalities, governmental agencies, and other entities that "provide" online services. To receive the protection of this federal statute, an organization must register under the act with the U.S. Copyright Office and follow the DMCA' s removal provisions. Check out http://www.loc.gov/copyright for the details on the DMCA.

The United States was the first country to enact DMCA legislation, and this statute took it into compliance with an international copyright treaty (the WIPO Copyright Treaty). As other countries imitate this general approach, the ability to cause infringing material to be removed without using expensive litigation will begin to become globally codified. See Julia A. Gladstone's chapter in this encyclopedia, "International Cyberlaw," for more on these developments.

Another significant U.S. legislative enactment was the No Electronic Theft Act of 1997 (NET). Under the NET, criminal penalties can be imposed on people who exchange or barter unauthorized copies of software, videos, clips, or music, whether or not they receive money for it. The only requirement is that the value of the pirated material exceed $2,500 (for felonies) in a given 6-month period. Although enforcement of the NET Act has been limited to high-profile cases thus far, all users should be aware of its provisions.

Among other important copyright areas, cases have held that publishers must pay additional fees for pre-Net work by freelancers, including a U.S. Court of Appeals ruling that the National Geographic Society made an unauthorized use of pictures taken by a freelance photographer back in 1961 when it issued a CD-ROM years later of its back issues (*Greenberg v. National Geographic Society,* 2001). It was ordered to pay license fees for that use. The U.S. Supreme Court decided the issue when it later ruled that media companies must obtain the consent of their freelance writers and creators (as employees create "works for hire," their employers typically gain those copyrights) before any pre-Net text, picture, or creation could be posted or sold online, thus forcing royalties to be paid for that use (*New York Times v. Tasini,* 2001).

In addition to the above, other developments occurred pertaining to the online copyright infringement issue. The U.S. Supreme Court held in *Eldred v. Ashcroft* (2003) that Congress had acted constitutionally in 1998 when it extended copyright protection for most works through the Sonny Bono Copyright Term Extension Act (Supp. 1999), retroactively increasing the copyright protection term from 50 years after an author's life to 70 years. Various cases upheld the constitutionality of the DMCA, including its safe-harbor provisions for ISPs. An issue being court-answered is whether the DMCA mandates ISPs to turn over customer information to the recording industry of those suspected of illegally trading music files; in *Recording Industry Association of America v. Verizon Internet Services* (2003), a federal judge ordered the ISP to identify an Internet customer who used music-file-sharing services, and this case was promptly appealed.

Whether the situation concerns Napster imitators, video downloading, or copying overseas, the legal wrangling between the copyright holders and the public over "fair use" will continue unabated. As other countries have or enact their own cyber copyright and intellectual property laws, the likelihood of global jurisdictional disputes will also increase.

DOMAIN NAMES AND TRADEMARK LAW

Trademark Law and Domain Names

For decades before the emergence of the Internet, trademark law was relatively straightforward. Trademarks and service marks abound, simply arising from a company's use of marks identifying certain products or services as being theirs (e.g., Apple with its rainbow apple and eBay's stylized logo). The concept of trade and service marks came about to keep businesses from "passing off" their products as being those of their competitors or of rightful owners, and the ownership of marks is an important intangible property right, along with copyrights, patents, and trade secrets.

Then along came the Internet, domain names, and new Web sites that had registered the addresses of bona fide trademark and service-mark holders (i.e., Burger King versus the "burgerking.com" registered by an individual). They knew that the domain name used to access any site was an important asset of identification, just as an entity's mark was important in purchasing a product (or domain name selection). As domain names must be registered to have any validity and registrars don't conduct background checks over an applicant's representations as to who owns the legitimate trademark rights, registration even today can be a race to the swiftest on a "first come, first serve" basis.

In the early and mid-1990s, "entrepreneurs" recognized this grand opportunity. In a style reminiscent of the old Gold Rush days, they raced to tie up as many of those good corporate names as they could, whether it was "harvard.net" or "burgerking.com." Later, they would send a demand letter to those entities and offer to sell their domain names back—at a tidy profit. Or the new owners would sit on their names and wait for that interest, conjuring up the concepts of "cybersquatting" and "cyberpirating." When one adds to this equation the various classifications (i.e., from ".com" and ".org" to the later introduced ".biz" and ".info") with the different country designations that are possible, it is easy to see the large opportunities created to tie up good corporate names at a good profit.

A brisk market in the buying and selling of domain names began—just hit the key word "domain name" in your search engines and see what arises. The people who registered general names, such as "business.com" or "loans.com," made excellent business decisions. One Houston businessman paid $150,000 in 1997 for the rights to "business.com," then sold that to a California company for a cool $7.5 million 2 years later. In 2000, mortgage.com was sold for $1.8 million and loans.com for $3 million.

Without any statutory guidance, the courts handed down mixed decisions as to when a mark holder would prevail, if at all, over a cybersquatter and any given domain name. The reason: The law was clear at the time that domain-name registrations and trademarks and service marks, whether registered or not, were two different concepts. Because there was no right by itself to use a mark as a domain name, owning one didn't necessarily convey ownership rights to the other.

Anticybersquatting Consumer Protection Act

In response, the United States passed its Anticybersquatting Consumer Protection Act (ACPA) in late 1999. This act allows civil lawsuits to be brought for trademark and service mark violations against anyone, who with a "bad faith" intent to profit from a mark, registers, uses, or attempts to sell a domain name that's identical or confusingly similar to that protected mark. Factors indicating bad faith are whether the name owner actually diverted the trademark owner's customers, offered the registered name for sale without having used it, or registered multiple names. Under the Anticybersquatting Act, the courts can cancel a "pirated" domain name, assess attorney fees and costs, and levy penalties of up to $100,000 against an infringer (depending on the level of bad faith and the actual damages).

This legislation also made it illegal to register the name of any living person without that person's consent, while intending to profit by that action. Actors Brad Pitt and Kenny Rogers immediately filed suit on this provision alone, Kenny Rogers objecting to the "kennyrogers.com" registered to the Web site of a California wedding service. Both celebrities retrieved their "names" back.

Late in 1999, the Internet Corporation for Assigned Names and Numbers (ICANN) decided that domain name registrations had to be prepaid. Before this change, cybersquatters could "hold" names for speculation; if they didn't find a buyer, they let the unsold names expire for nonpayment of the registration fees. The amount of statutory damages for copyright infringements was increased. An amendment to existing trademark legislation was enacted, permitting trademark dilution grounds to be used as a cause of action in canceling domain-name registrations. In late 1999, ICANN then announced that more than 800 addresses of "dash dot" domain names had been revoked. A software glitch apparently had allowed these registrations to proceed, and the excesses of the past were being corrected.

ICANN's Dispute Resolution Process

ICANN next instituted a procedure for resolving domain-name disputes. Under an agreement called the Uniform Domain-Name Dispute-Resolution Policy (UDRP), those with a dispute over a registered domain name involving their trademark or service mark have the alternative to file a complaint with an ICANN approved dispute-resolution service. Pursuant to these procedures, the service provides an arbitration panel that then rules which party has the legitimate right to that name, and if either party disputes the ruling, then that party can litigate the matter further in court. The judge in the court case can review all of the facts and isn't necessarily bound by the review board's determination. Once a final decision is reached, the registrar transfers the domain name as the court or administrative panel decided.

Whether marks are registered or not, legitimate holders can take their case to different alternative dispute resolution centers under this process, and the United Nation's World Intellectual Property Organization (WIPO) hears most of these cases (http://www.wipo.org). Basically, the claimant must prove (a) that the domain name very closely resembles a trademark registered or owned by that entity, (b) that the party that registered the domain name has no rights or legitimate interest in that name, and (c) that the domain name was registered and used for illicit purposes or in bad faith. This is an inexpensive process (a one-person panel at the time of this publication costs a complainant $1,500, and a three-person panel costs $4,000); fast (the arbitrators' decision is rendered within 45 days); convenient (there is no hearing to attend; the arbitrator or panel reviews only the complaint, response, and supporting documents); and the decisions typically favor the trademark holder (some four fifths of the determinations favor the mark owner). For further details on ICANN or its UDRP procedure, see http://icann.org.

The initial decisions on domain names reached conclusions in favor of companies with recognizable names, such as the World Wrestling Federation, Stella D'Oro Biscuits (a Nabisco affiliate), and Telstra (the Australian telecom company), ordering their domain names transferred back to them. WIPO arbitration panels handed back the Web addresses bearing the names of the Corinthians (Brazilian soccer team), Dan Marino, Julia Roberts, Kevin Spacey, Yahoo!, ESPN, and Wal-Mart. However, Sting, Bruce Springsteen, the Reverend Dr. Jerry Falwell, and Ted Turner failed to prove that their personal names had been used in a trademark sense as a label of particular goods or services and did not prevail in their UDRP proceedings. ICANN's dispute-resolution policy, however, does not apply to all registrars—only to the TLDs (top level domains of ".com," ".net," ".biz," etc.) and not to the "ccTLDs" (individual country domains, such as ".cn" for China), unless that country agrees or has established its own dispute resolution board.

The Legal Weapon Tradeoffs

Entities with U.S.-based domain name/trademark conflicts must decide whether to use the Anticybersquatting Act, an ICANN proceeding, or both. ICANN gives a fast resolution, whereas ACPA litigation can take up to 2 years or more for a decision. Although the federal statutory-authorized lawsuit allows for a preliminary injunction, the opportunity to be awarded good damages, and transfer back of the domain name (which is all that an ICANN-UDRP procedure can do), this alternative is highly expensive with much more downtime, complexity, and legal dollars required. The ACPA is a final determination, whereas the extent to which an ICANN decision can be litigated further is currently being determined.

A federal appeals court in *Sallen v. Corinthians Licenciamentos Ltda*. (2001), however, has ruled that the ACPA granted U.S. federal courts the jurisdiction to review and even override findings or decisions made by a dispute-resolution service provider under an ICANN proceeding. Lower U.S. District Court decisions have upheld ICANN-UDRP's validity and the right of courts to review and even overturn those decisions (see, for example, *Weber-Stephen v. Armitage,* 2000; *Parisi v. Netlearning,* 2001; and *Virtuality v. Bata,* 2001). One U.S. Court of Appeals decision in *Mattel v. Barbie-club.com* (2002) has held that in rem proceedings (deciding ownership rights to property, such as domain names) under ACPA may only be brought in the judicial district where the domain-name registrar is located—a large problem with international disputes or where multiple domain-name owners are involved. Owing to the negative aspects of the ACPA litigation alternative, the great majority of entities at this time do choose to employ only an ICANN proceeding.

The rest of the world is not as restrictive as the United States or other developed countries, and the game is still being played in some fashion. Countries in Asia, the former Soviet Union, Eastern Europe, and others have not yet tightened their laws, although over time, they likely will. It may still be possible in small nations such as Moldavia (until they change) for cybersquatters to purchase domain names that aren't already reserved, and the question of conflicting laws always seems to rear up when lawsuits are brought to challenge ICANN administrative rulings.

For international protection, entities need to register their mark with the U.S. Patent and Trademark Office (PTO), and U.S. concerns need to file that registration with the appropriate agency of the foreign countries in which they operate. Although the granting of a registration by the PTO is not conclusive on the issue of who owns a mark, it is prima facie evidence of that ownership (see the PTO at http://www.uspto.gov for the details).

Given the estimated numbers of domain names (more than doubling from present numbers to more than 100 million in 3 years by various estimates), the number of permutations, and various country designations, the volumes of conflicts, arguments, and opportunities for these disputes can only increase over time.

PATENTS

Through patents, the United States by its Patent and Trademark Office grants an inventor the exclusive right to make, use, or sell an invention for 14 years (design patents) or 20 years (inventions). To obtain a patent, the inventor must meet certain requirements, such as novelty, usefulness, and nonobviousness. Computer hardware (i.e., the design of electronic components, handwriting recognition systems, and so on) is patentable, but software typically isn't brought into the process, because of the time it takes for a patent to be granted and contrary regulations in a few cases. The number of patent applications for Internet applications increased in recent years owing to court decisions (beginning with *State Street Bank v. Signature Financial Group,* 1998) upholding patents issued for Web site business processes and operating methods. Amazon.com (its "one-click" online purchasing system), Onsale (operating auctions), Cybersettle (its "double-blind bid" dispute resolution procedure), Priceline.com (conducting an online "reverse" auction), Microsoft (online shopping and merchandising), Sun Microsystems (Internet bill processing), Tumbleweed Communications (online greeting card delivery), E-Data Corporation (online selling to any point of sale location), and other e-commerce entities have received patents on their specific business or operational models. As expected, the court challenges from competitors over these patents have also increased, as competing entities battle over controlling important operational methods. In this encyclopedia, Gerry Bluhm's chapter on patent law details this subject.

DEFAMATION

Cyberspace creates an inordinate ability to quickly post defamatory comments that injure another person's reputation and can be seen around the world. Those who post libelous comments, however, do not enjoy the anonymity in cyberspace that one might expect. A cyber-defamation lawsuit is typically filed in the city where the chatroom provider or ISP is located and lists various "John Does" or "Jane Does." This is legal jargon for naming unknown parties to the lawsuit, whose actual names will be added later on after discovering their true identities. The lawyer then files a subpoena (demanding that the desired information be released) against the provider to gain the identity of the particular John or Jane Doe who posted the inflammatory remarks.

Lawyers serve subpoenas daily on CompuServe, AOL, Yahoo, Microsoft, and others to retrieve some poster's identity. If the ISP doesn't turn over the demanded information, the attorney then goes to court to ask the judge to force the ISP into divulging the required data. The judge balances the right to protect someone's anonymity versus the injured party's right to be protected from harm. The plaintiff, or injured party, usually must prove that there's no other way to obtain relief without securing this specific information. If multiple servers are involved, the attorneys will follow the e-mail address back through that chain with multiple subpoenas.

One doctor at the Emory University School of Medicine in Atlanta, Georgia, came across a posting on a Yahoo message board. It falsely suggested that he had taken kickbacks from a urology company to give his department's pathology business to the company and had been forced to resign over this conflict of interest. The message was from "fbiinformant," who later was discovered to be a former employee at the urology company who disliked the doctor. In what was believed to be the first Internet defamation case to reach trial, a U.S. District Court judge awarded $675,000 to that doctor, all because of this one "anonymous" Internet message (see *Graham v. Oppenheimer,* 2000). Litigating defamatory e-mails, postings, and communications continued on unabated from there.

As to the online service providers, the court decisions have consistently upheld that there is no liability on their part for defamatory postings made by third parties. The Communications Decency Act of 1996 (CDA) bars tort-based claims or lawsuits against ISPs for defamatory, obscene, or other objectionable postings, provided there is no complicity by the ISP with that third party over those postings or other unreasonable behavior (the provider also must "actively" remove the objectionable material). Although portions of the CDA were later held to be unconstitutional (see *Reno v. American Civil Liberties Union,* 1997), these provisions continued in their legal validity and effect.

PRIVACY CONCERNS

At this time, there is no general, sweeping U.S. law regulating or requiring entities to disclose how they use sensitive financial and other personal information gained from their customers or users, nor how they gather that data. Notwithstanding recent U.S. Federal Trade Commission (FTC) high-profile proceedings against companies over their information practices, the FTC has had, in effect, a "self-policing" policy, promoting industry self-regulation in the fields of data collection and customer profiling. Congress did enact legislation as to the online privacy protection of children, however, when it enacted the Children's Online Privacy Protection Act (COPPA) in 1998. This statute requires that Web sites "earmarked" for children, or who knowingly collect data on users under age 13, need to (a) obtain verifiable parental consent for any collection or use of their children's information (i.e., "no consent, no collection") and (b) on request, provide the parent with the ability to review any personal information that has been so collected. The FTC has issued administrative rules on COPPA to guide these "kiddie" Web sites on their compliance with this legislation (see http://www.ftc.gov for more on this subject). Additionally, the FTC has charged entities with violating COPPA and reached settlement with several of those Web sites.

Congress also passed the Gramm–Leach–Bliley Act (GLBA), known also as the Financial Services Modernization Act of 1999. This legislation requires that financial institutions (a) inform consumers of their privacy policies and (b) notify consumers before acquiring, transferring, or selling their private data to third parties, giving them the opportunity to "opt out" of such data-transfer practices. Although the GLBA applies to the Internet and use of cookies, it remains to be seen how these basic provisions will be enforced as a whole.

Additionally, various states have passed privacy legislation both in the financial area and in preserving the confidentiality of sensitive medical records, including diagnosis, treatment, and prognosis. As with most state laws, these acts are not uniform—for example, depending on the subject area, some states require opt-in conditions before data can be collected or used, whereas others establish opt-out standards. Rules have also been enacted under the Health Insurance Portability and Accountability Act (HIPAA) of 1996 that deal in this area (see the U.S. Dept. of Health and Human Services at http://www.os.dhhs.gov for more information).

The differences between the United States and other countries, such as Canada and the European Union (EU), couldn't be more pronounced than in the rights of privacy area. Canada's Personal Information Protection and Electronic Documents Act (S.C., 2000) implements a wide array of data protection, including a phase-in of "opt-in" provisions for its citizens. The EU's Data Protection Act came into effect in 1998, also using a completely different approach from that of the U.S. Citizens there gained enhanced rights to prohibit their private data from being released, including requiring entities to secure "opt-in" permission in various situations before personal data can be acquired, sold, or shared. Although it remains to be seen how these provisions stand over time after later court reviews, statutory additions, and individual member country changes, these policies are more supportive of their citizens' privacy than those of the Unites States. Additionally, pursuant to the EU's Directive on Data Protection, EU countries may block the transfer of their citizens' personal data to countries that do not guarantee adequate privacy protection equal to theirs. Although the United States and EU negotiated "safe harbor" guidelines for U.S. subsidiaries that operate there (basically, the U.S. company must certify to the U.S. Department of Commerce that it is in compliance with the EU's tougher laws), global policy differences such as those of the EU and other countries place pressure on the United States to adopt generally more consumer-sensitive privacy policies. For further information, see the U.S. Department of Commerce Web site at http://www.doc.gov, as well as Pfaffenberger's treatment of these considerations in his chapter of this encyclopedia.

CENSORSHIP

The First Amendment places strong limitations on the government's ability to censor, or unduly regulate, the rights to basic freedoms such as those of speech and expression, and the Internet is no exception. For example, Congress passed the Communications Decency Act of 1996 (CDA) which, among various provisions, essentially made it a crime for anyone to knowingly distribute obscene material for sale in cyberspace. Later in 1997, the U.S. Supreme Court in *Reno v. American Civil Liberties Union* declared that most of the important provisions of the CDA dealing with obscenity were unconstitutional, holding that these provisions were so vague as to be void on their face. The fact that statutes applied to cyberspace didn't mean the constitutional tests applied were any less strict, and later legal battles over subsequent pornographic statutes have applied the same strict construction tests of the offline world.

For example, Congress in 1998 passed the Child Online Protection Act (COPA) as a successor to the struck-down CDA provisions in another attempt to stop children from gaining access to sexual materials on the Internet. In 2002, the U.S. Supreme Court reviewed COPA and partially upheld it, adding further legal uncertainty. In *Ashcroft v. ACLU* (2002), the court ruled that the act's reliance on community standards to identify "material that is harmful to minors" did not by itself render the statute unconstitutional. The divided court kept alive the fight over whether the measure was unconstitutional, however, by affirming the appellate court's ban on COPA's enforcement, then returning the issue to a lower court for further review or trial on free-speech questions the court felt were still unresolved. Legal experts feel that COPA runs into legal problems similar to those faced before by the CDA.

A three-judge panel ruled that the U.S. government cannot legally require public libraries, as a condition of receiving federal funds, to install mandatory filters on every computer to block online children's pornography. Accordingly, the panel struck down the Children's Internet Protection Act (CIPA, 2000), holding that the filter requirement blocked constitutionally protected speech and that there were less restrictive alternatives available (i.e., establishing "acceptable use" policies, requiring parental consent, and placing unfiltered machines away from public view). The two cases, *American Libraries Association v. U.S.* (2002) and *Multnomah County Public Library v. U.S.* (2002), were appealed to the U.S. Supreme Court.

The Supreme Court also granted review to the "Nuremberg Files" case, in which an 11-judge U.S. Court of Appeals panel voted 6–5 in *Planned Parenthood v. American Coalition of Life Activists* (2002) to uphold a 1999 trial verdict against an anti-abortion Web site that disclosed the names, addresses, and family information of physicians who performed abortions. The trial jury awarded punitive damages of $107 million against this coalition of activists, but the appellate panel remanded the damage amount back to the trial court to determine whether the award was excessive or not.

In a case in which the judge described was a dispute between the protection of trade secrets versus freedom of speech, a Federal judge ruled that a Web site operator could continue posting confidential Ford Motor Company documents on new car designs and other internal data (see *Ford Motor Co. v. Lane,* 1999). Lawsuits over employer Internet-use policies, however, provided the policies are reasonable with notice given to the employees, have generally been upheld in the employer's favor. An employer's use of Web filtering software (where worker e-mails are subject to scrutiny) also is under challenge, but most legal scholars believe that these challenges generally won't be sustained. Moreover, the courts so far have not been overly receptive to employee claims that work done on an employer's computer system is personal and entitled to constitutional protection.

Courts have generally upheld the right of "suck.com" Web sites to criticize the operations of major businesses (e.g., "http://www.ballysucks.net"; see *Bally v. Faber,* 1998), although the complaining companies have had more luck in challenging these sites under ICANN domain name proceedings. For example, the Salvation Army filed an ICANN-UDRP proceeding over the name "salvationarmysucks.com", and the arbitrators ruled in favor of it, when evidence of offers to sell the name back to the Salvation Army came to light. Generally, provided the posted allegations are basically true or represent protectable opinion rather than false facts, the cases hold these expressions to be protected by the First Amendment. If you use a search engine with the descriptive word "suck.com," surf to the site of "Sucks.com," or type your favorite company name with a "suck.com" after it, watch the complaint Web sites surface that already are in existence.

Another growing First Amendment consideration involves its use in controversies in which software sites collide with trade interests. For example, the DVD Copy Control Association (DVDCCA) attempted to shut down Norwegian teenager Andrew Bunner's Web site, which provided a link to software code that unscrambled encrypted DVDs. A California Court of Appeals court (*DVDCCA v. Bunner,* 2001) overturned the trial judge's temporary injunction against Bunner's site as an unconstitutional prior restraint, citing that the First Amendment protected his publication of the DeCSS software as an exercise of free speech, even given a continuing violation of DVDCCA's trade secret rights. The California Supreme Court currently is hearing the appeal; once this issue is decided, the main case on whether the defendant misappropriated trade secrets or committed copyright infringement by the DeCSS posting is to be decided, along with whether a permanent injunction or damages would be appropriate. For more on the subject of censorship, please see Julia A. Hersberger's chapter.

CYBERFRAUD

Cyberfraud is a major problem on the Internet, owing to its anonymity, commercial reach, and speed in exchanging credit card information. The ease of entry onto the Internet makes it even easier for "fly by night" firms to race in, skim off money, and quickly disappear without leaving a trace. No matter where you live, the Web has indeed become global in when and how the unwary are fleeced. In the United States, the FTC is the federal agency charged with prohibiting unfair or deceptive commercial acts, including misleading advertising (fraudulent investments are handled by the Securities and Exchange Commission, or SEC). The FTC's rules and regulations against unfair or deceptive business practices specifically cover Internet transactions, and its Web site (http://www.ftc.gov) has pages with information about Internet fraud, complaints, and its fight against this problem.

The FTC has brought countless numbers of fraud and misleading advertising complaints to stop such unfair practices, both online and offline. Sold ever increasingly over the Net, the first batch of the FTC's "Top 10" online scams centered in Internet auctions, ISP fee promotions, Web site design/promotions and cramming (being billed for services or products not authorized), Internet credit card cramming, multilevel marketing/pyramid schemes, business opportunities (i.e., software and sales motivational tapes) and work-at-home pitches, investment schemes (i.e., day trading, gold deals, and oil and gas leases) and get-rich-quick scams, travel/vacation fraud, telephone/pay-per-call solicitation, and healthcare scams. From year to year, identity fraud is the overall top consumer fraud complaint from all sources received by the FTC by a wide margin (over 40% of all complaints received on its Consumer Sentinel database), followed by Internet auctions and Internet services/computer complaints.

In 2001, the FTC opened up a Web site to the Net public with statistics and information on primarily U.S. Internet fraud and identity theft. The site is called "Consumer Sentinel" (see http://www.consumer.gov/sentinel), and it started with a database of more than 300,000 complaints lodged with the FTC over the last several years. The site provides data on fraudulent transaction trends, as well as the ability to submit online complaints (as does the FTC with its Web site). If you find Internet fraud problems involving different countries, then head to "econsumer" at http://econsumer.gov, which has in place direct links to consumer fraud agencies in numbers of countries and represents a coordinated world effort to work together on this global problem. This site also allows for the filing of international fraud complaints.

The Internet is both an excellent tool for investors with easy access to research sources, as well as for the shysters, hipsters, and con artists to that money. Because anyone with a computer and a modem can reach tens of thousands of potential investors simply by creating an attractive Web site and spamming, the U.S. SEC has had to become very active on the Internet. It established a national cyberforce of attorneys, accountants, and analysts specifically trained to watch out for fraudulent online security transactions (see http://www.sec.gov for more information, including its reported Internet fraud cases).

For example, one case involved a company that used spam e-mail to announce an initial public offering (nearly all of us have received junk e-mail of this type), stating that it had been approved by the SEC and would realize at least $1 billion in eyewear sales. In fact, it had never been approved by the SEC and didn't even own an office or inventory. The owner of the company used the "invested" money for restaurants (eating meals), casinos (gambling), and adult entertainment clubs. The investors lost everything that they had invested—the usual and unfortunate result with fraudulent investments, regardless of whether the FTC or SEC is able to intervene.

Regardless of the state or country, the law is clear with respect to cyberfraud: (a) any Web advertisement of illegal transactions under a particular country's or state's laws (e.g., gambling or usury) will be illegal there; (b) fraudulent, false, or misleading statements are illegal and unenforceable, no matter where you live; and (c) the regulatory agencies in different states and countries vary widely in their ability to crack down on misleading advertising, even when the customers or investors have lost all of their money. Remember: if the advertised "return" is too good to be true, then it usually is, and investors must be quite careful when reviewing potential investments of any kind, whether online or offline. (For more information, see Marc D. Goodman's chapter on cybercrime and fraud.)

E-COMMERCE LAW

Given the ease with which online contracts can be made, the tearing down of geographic barriers, and E-mail "proof" that lasts forever, basic contract law is even more important on the Net. From what is needed for a legitimate contract (i.e., mutual assent, consideration, capacity, and no legal defenses) to how duties are delegated and determining damages, all of the fundamental contract laws apply in cyberspace. (For an excellent discussion on this area, see *Cyberlaw, Text and Cases,* by Ferrera, Lichtenstein, Reder, August, & Schiano, 2001). As expected, however, e-contracts do have important aspects that stand out from offline contracts and as discussed in the next section.

"Click" Contracts

You can't sign an e-contract with your real name, as you could if you were in a face-to-face meeting with a sales rep. In place of this, the law basically allows that you can agree to the terms and conditions of an electronic agreement when you click the "I agree," "I'll buy," or "Subscribe" button. The mouse click on that agreement button sets the approval to the conditions of the e-contract.

The courts generally have upheld this as being as valid as if you signed a written contract on the dotted line. (See, for example, *Crispi v. The Microsoft Network,* 1999, and *Geoff v. AOL, Inc.,* 1998). Any on-screen click, no matter where it's located (but provided it's reasonably identified as the "click" agreement button), will do. The legal premise is that the medium in which a signature or contract is created shouldn't affect its validity, and the transaction is enforceable, whether that medium is paper or electronic. There are limitations on when click contracts will be enforced, however, and this is discussed at the end of this section.

E-signatures

The U.S. government and nearly all of the states have enacted some version of an electronic signature statute. The federal act (Electronic Signature in Global and National Commerce Act, 2000) ensures that electronic records, signatures, and contracts have the same legal effect as their ink-and-paper counterparts (including that electronic records satisfy statutes mandating that records be kept in writing), validating online commerce and allowing for the eventual recordation of documents such as deeds, mortgages, and bills of sale by accepting electronic notarization. This legislation (including state counterparts) mandates that an e-signature is enforceable if both parties agree to its technological format, and it provides that a signature may not be denied legal effect simply because it is in electronic form. Signature verification can be based on fingerprint scans, smart cards, encryption software, mouse-pad signatures, eye lasers, voice recognition, or whatever the parties agree to use in their contract. It is an enabling statute that sets down standards that can be followed and allows states to enact their own but generally consistent legislation within this umbrella, prohibiting laws that limit permissible electronic signatures to one single technology. The states differ widely in their authorizations because numbers are applicable to all electronic transactions, whereas others can be more limited in their scope and effect. Internationally, many countries (from Brazil and Taiwan to the individual members of the EU) have enacted e-signature and e-commerce laws, simply because it is in their best interests to do so.

Taxation

In 1998, the United States' Internet Tax Freedom Act capped taxes on online sales with a 3-year moratorium on any new state, local, or federal Internet taxes (as defined) to October 20, 2001; after an interim delay, the moratorium was extended for an additional 2-year period that expires on November 1, 2003. It is a misnomer, however, that this act generally ended taxation of the

Internet. Among other allowed taxes (including "grandfathered" Internet taxes and taxes on Internet access) when the "moratorium" expired, 45 states charged sales tax on tangible products brought online, given that the seller had some form of a physical presence or "nexus" (i.e., a warehouse, retail store, office, or sales representatives) in that state. Under existing tax law (see *National Bellas Hess v. State of Illinois,* 1976), mail-order houses and, in turn, Web sites can collect sales and excise taxes on the sales of goods when such a connection or nexus was present with the taxing state. What was held in abeyance was the legal ability to tax an online purchase from a resident in a state where no such nexus or connection to the selling site's state was present.

The position of the World Trade Organization and international agreements generally parallel the United States' tax moratorium and approach. The pressures not to follow this are strong, however; as one example, the EU already taxes certain online sales of goods such as books and CDs. As the Internet erodes into the base of revenues that governments receive from traditional "brick and mortar" stores and their sales, it is likely this taxation will increase over time—the only question is when and how. The first "opening" is after November 1, 2003. Although Congress could again extend the tax moratorium, it is expected that legislators will listen to the pleas from the states and non-Internet merchants to tax online sales as never before and the constant lobbying will continue well past that date. See William A. Duncan's detailed chapter on taxation for more.

Spam

As e-commerce expanded, huge increases in unwanted electronic solicitations (otherwise known as spam and junk e-mail) also occurred. Although more than 20 states in the United States have enacted laws now regulating the distribution of commercial bulk e-mail (including fines and jail time as penalties), these laws are rarely or difficult to be enforced. In some states, no one has yet to be prosecuted under those statutes. Part of the problem lies in the Net's anonymity, enforcement difficulties, and jurisdictional issues, not to mention that more serious crimes are usually in line for attorney generals and district attorneys to prosecute. Until a federal antispam law in support of the states becomes effective with a strong enforcement mechanism—not to mention such laws also being implemented in other major industrial countries—controlling spam legally will not be possible. For the latest on antispam legislation and developments, see http://www.spamlaws.com.

International Aspects

Although beyond the scope of this chapter, the global e-commerce aspects of cyberlaw are considerable. Whether the subject involves the EU's passage of e-signature, copyright, privacy, or e-contract laws, or the cyberlaws of major industrial countries such as Japan and Germany, these international aspects are continually evolving. Furthermore, the treaties that can affect cyberspace range from the General Agreement of Tariffs and Trade (GATT) to the United Nations Commission on International Trade Law (UNCITRAL) and beyond. Please see the treatment of these international areas throughout this encyclopedia.

E-COMMERCE "TERMS OF USE" PROVISIONS

Many Web sites separate out the necessary purchasing information from their legal Terms of Use and Privacy policies. Although the general areas treated are quite similar, each site's specific provisions are different, depending on whether a particular location sells products, provides services, gives information, or some combination. Nonetheless, the general concepts covered are basically the same.

Given that the basic contract business terms (e.g., quantity, price, time for delivery, delivery mode, and so on) are present, what was once "just legal boilerplate" has now become more important: disclaimers of liability, indemnity, handling of disputes, applicable law, and dispute resolution, among other areas—and this is especially true when distant localities become involved. The following "boilerplate" provisions are typically found in the "Legal Provisions" or "Terms of Use" sections:

Disclaimers of Liability

This section typically limits a seller's liability for injury or loss incurred by the buyer to exchanging the product or a refund of the purchase price, all at the seller's option. This refund policy is usually coupled with a disclaimer of liability, such as the following: "Seller disclaims all liabilities and warranties, express or implied, including the warranties of merchantability and fitness for a particular purpose, and this Web company shall not be liable for any damages, whether consequential or incidental."

The intent of these provisions is for the seller to "duck away" from liability, leaving the manufacturer on the hook. Although not unanimous in their decisions, the courts tend to uphold limitations on consequential damages (for example, loss of data that results in a further loss of customers or business) in commercial transactions. They usually are not upheld if a personal injury is involved, however, such as when a customer becomes injured when using a defective product that caused those injuries.

As a basic legal concept, no one can contract against the effects of strict product liability or their own negligence. If that were the case, there could never be any product liability, because every manufacturer and retailer in the world would be contractually providing, "Sorry, if there's a problem with our product, even if it's all our fault, we don't accept that liability—you do."

Depending on the circumstances, Web sites can be held completely responsible for a customer's damages, notwithstanding these one-sided contract provisions. For example, if a pharmacy Web site erroneously fills a prescription for high-blood pressure medication that severely worsens the problem, that Net retailer typically will have to compensate that injured person for his or her damages from those bad pills, regardless of any Terms of Use limitations to the contrary.

Indemnity

Standard Terms of Use agreements also provide "tight" indemnity provisions. These terms basically provide that the buyer or user indemnifies, or is responsible for, any damages that occur to the Web site by reason of how the buyer uses that product, but do not place the same indemnification responsibility on the Web site toward the buyer. Although courts tend to uphold these terms, they will look at how "fair" the provisions are, and some have been thrown out; this result does depend on the facts and applicable state law in any given case.

Handling of Disputes

Captions such as "Dispute Resolution," "Conflicts," or sometimes simply "Legal Fees" usually head this section. A growing trend is for Web sites to contractually require arbitration and mediation as the agreed methods to handle disputes—alternatives that are cheaper, faster, and more convenient than traditional litigation. (Cyberlaw Dispute Resolution is treated later in this chapter.) Should litigation flare up, the typical section also authorizes the payment of attorney fees and court costs to the "prevailing party," or the one that wins in court.

Applicable Law, Dispute Location, and Other Provisions

This is another important area, covering what law is to be applied and where disputes are to be heard, with Web sites generally choosing their hometown and those laws. If that law isn't entirely in the site's favor, it will always designate a place that is. Just because a site specifies that a certain state's or country's laws or situs applies doesn't necessarily mean that a court will uphold that designation (see "The Clash of Laws" for this discussion).

Terms of Use agreements contain various other provisions. For example, they may provide for copyright and trademark notice requirements, restrictions on chat-room conduct, forum rules, prohibitions on posting defamatory comments, linking conditions (e.g., "you need our permission to link here"), acquire the copyright to any postings made, rules for games or sweepstakes, address for sending notices, and how notification is to be made, among other provisions.

Web operators typically place a copyright notice on their sites (for notice purposes, even if it's not legally required) and limit what can or can't be done when you copy from their sites, whether this limitation is legally correct or not. For example, "You may not reproduce, post, transmit, or distribute in any way whatsoever any information or display from our site, regardless of its source, without our prior written permission." Regardless what is posted, Web sites cannot limit what the law already allows. For example, the United States provides for a "fair use" exception for copyrights, and the courts will not allow any such one-sided limiting or restrictive statements to erode this long-standing doctrine of "fair use."

Privacy Policies

The surveys conclude that to maximize customer "hits" and use, a Web site must provide reasonable, detailed, and understandable statements on how it protects a user's privacy and then follow these policies without fail. Net customers are skittish as to the sanctity of their sensitive data—and they should be, given the pace at which today's technological advancements are eroding their comfort zones.

Privacy policies do not necessarily have to be long-winded, detailed, or complex—and they shouldn't be. See the previous section, "Privacy," and other chapters in this encyclopedia for more on this topic.

Validity

Most courts uphold "take it or leave it," click e-agreements, provided (a) the terms are written in understandable English with readable print and not hidden from the user's view; (b) the user has the opportunity to read and understand these terms, all before having to make any purchase or use decision; (c) the provisions are reasonable; and (d) the user has to take some affirmative action to agree (such as clicking an "I agree" button). As courts tend to enforce software shrinkwrap agreements, they're doing the same with Web-site Terms of Use provisions, provided these elements are present.

If, however, the language used is hidden or not conspicuous, in small print, or wholly unreasonable in effect (e.g., "This Web site is not responsible for any of your damages, regardless of how much we are at fault"), then courts will generally not uphold them. There must be some knowledge (the terms are easily located and understood), prior decision making (the user can decide before having to order), and facts showing at least mutual assent or an implied agreement (e.g., clicking on an icon to show your assent).

A leading Second U.S. Circuit Court of Appeals case, *Specht v. Netscape Communications Corp*. (2002), reviewed the standard terms and provisions applied when a user downloaded software. The standard terms provided that all provisions came into legal effect when any user simply downloaded or installed the software. However, users could directly download this software before coming or scrolling down to the "I agree" icon and any inspection of the terms of use that bound their decision. Applying standard contract law on mutuality of assent, the court upheld a lower court's decision and refused to enforce an arbitration clause in the Web site's forum state, holding that the required mutuality for contract assent was not present.

When the requisite criteria is present, courts will uphold these terms on a general, conceptual basis, provided additionally there is no blatant unreasonableness in those provisions. Based upon the Uniform Commercial Code's Section 2-302 (basic to most states' laws) that codifies the traditional common law doctrine of unconscionability, courts still have the ability to set aside those standard terms. The intentional misuse of a Web site by a user, however, can negate provisions that don't meet these standards. For example, in *Register.Com, Inc. v. Verio* (2000), a court held that where Internet users intentionally misuse a site's content, there is an implicit agreement to any terms of the Web site's legal notice that prohibit such action.

THE CLASH OF LAWS

Given the numbers of cyberlaw conflicts, a key issue is whether a court in the plaintiff's geographical area can hear and decide the case. Simply put, if the problem is big enough, you want your understandable laws to apply, eat and sleep in your own home, work in your office when not in court, and not have to hire an expensive attorney in a different state or foreign country. People involved in a court case don't want jet lag, bad food, unfamiliar surroundings, a strange language, and separation from family for weeks on end—which is why jurisdiction and conflicts of law is such a large cyberlaw issue.

Jurisdiction over a nonresident Web site or Internet transaction in the United States is normally based on a local state's long-arm statute. These laws provide that a state can assert jurisdiction over a nonresident defendant who commits a tort, transacts business, or has some connection with that state. When a state court asserts jurisdiction over a particular controversy to render a binding decision, for the most part it will also be constitutionally permitted to apply its laws, given sufficient connections with that dispute. (Federal courts apply the appropriate state substantive law when relevant to their decisions).

The U.S. Supreme Court's landmark *International Shoe Company v. Washington State* (1945) decision established the law in this area. For a court to have proper jurisdiction, defendants must have purposely availed themselves of the privilege of doing business in that particular state (i.e., traveling in, selling there, advertising in, or other contacts in that state), and these "minimum contacts" must meet sufficient levels of due process so as not to offend our "traditional concepts of fair play and substantial justice."

For example, if a Wyoming rancher died and willed his ranch to his Wyoming son, it would offend our sense of "fair play" if a second son, who happened to live in Alabama, could sue and haul that Wyoming brother into an Alabama court. There are no connections with Alabama to this case, so the only court with jurisdiction should be Wyoming—and that court would apply Wyoming law (the property is there, the decedent and his heir lived in the state, the will was probated there, and so on).

When it comes to jurisdiction on the Net, U.S. courts generally look at a Web site's level of Internet activity, drawing distinctions between passive and active locations—and this distinction will be drawn more in foreign courts, although it is not yet a trend overseas. Called the "Zippo sliding scale," this test was set down in *Zippo Manufacturing Co. v. Zippo Dot Com, Inc.* (1997).

At one end of the spectrum, Web sites that enter into contracts with out-of-state residents involving repeated contacts, e-mails or other correspondence, and selling appreciable amounts of products or services into that state are held to be "active" Web sites. These active sites can be sued in the state of their customers, or in those out-of-state locations, as various U.S. cases have held. Given the existence of these sufficient contacts, a local court could (depending on the facts) disregard a contrary Terms of Use condition of the Web site that provided it could only be sued in its home state.

On the other side of the equation, sites that only advertise or post information about their business on the Internet—not taking any orders or conducting business through that Net site—are held to be "passive" Web sites. These Internet "informational" sites generally cannot be sued out of state and dragged from their home base into foreign courts, simply because they maintain a virtual presence. To hold differently would be to subject Net operators to being sued anywhere in the world that allowed an Internet connection to be made.

In between are Net sites that provide more connections between their state and out-of-state customers—and it usually takes more than having an e-mail capability and a toll-free number to confer that jurisdiction. E-mailing questions about a potential purchase (without more) doesn't suffice either, although some courts view this as borderline and "getting very close." If a defendant Web site displays a downloadable mail-in order form, toll-free telephone number, and e-mail address but no orders are ever made there, then these facts are not normally good enough. Given a finding that a passive site was involved, the upset user then must travel to the Web site's home state to sue. As indicated, a factual decision needs to be made in each case as to whether these minimum connections for due process reasons are present.

Some courts have developed an "effects"-based approach exception to the Zippo sliding scale. When using this approach, the court focuses its analysis on the actual effect that a Net site had in its state or the defendant's intent, not on how interactive the Internet location was. This test derives from the U.S. Supreme Court's *Calder v. Jones* decision (1984), where jurisdiction was found over a nonresident defendant newspaper, based on its intentional conduct outside the forum state that was deemed to cause defamatory injury to the plaintiff in that state—even though strong factual connections weren't otherwise present. Although courts tend to apply the "effects" test in cases involving intentional torts, such as defamation or trademark infringement cases, they do differ on what conduct constitutes the kind of "express" aiming or effect that's required to satisfy the test. Applying the "effects" test can find jurisdiction when there aren't sufficient contacts of the type called for under the *Zippo* sliding scale.

In a closely followed case, the California Supreme Court in *Pavlovich v. Superior Court* (2002) overruled a Court of Appeals decision involving the application of the effects test. Pavlovich had posted his programming adaptation of DeCSS software, a controversial technology allowing the scrambling system in DVDs to be rendered ineffective and their contents to be copied. The appellate court held that it had jurisdiction over Pavlovich, a Texas resident who was an engineering student at Purdue University in Indiana at the time of the alleged infringement. Although the defendant's actions didn't come close to meeting the *Zippo* test, the appeals court held that since there was an "effect" in California from Pavlovich's "intentional tortuous" actions, that California had jurisdiction, could apply its long-arm statute, and could force the defendant to come there and defend himself. The California Supreme Court, however, reversed this decision on a 4–3 vote. It held that under the *Zippo* test, the Web site was merely an informational one and there was no evidence that the defendant had expressly hurled his actions at California. It ruled that a defendant's knowledge or foreseeability alone of harmful effects ensuing in a specific state (California) is not sufficient to establish any "purposeful availment" of that state's law under the effects test. There must be more than that.

Keep in mind that the U.S. Supreme Court hasn't yet ruled on this area of virtual personal jurisdiction, nor what happens when a Web site's Terms of Use provisions are different from the reviewing court's laws—and courts, domestic and foreign alike, can and do give "wild" judgments that don't meet any tests or analysis. For example, Australia's highest court ruled in *Dow Jones v. Gutnick* (2002) that a publisher could be sued for defamation in whatever country an individual's reputation has allegedly been harmed. This case arose when an Australian businessman, Joseph Gutnick, sued U.S.-based Dow Jones & Co. for comments made about him in an article posted on the Internet in Barron's Online. This court found that since damage to the plaintiff's reputation had occurred in Victoria where the article was downloaded and read, it was appropriate for Gutnick to seek damages in that forum.

Because laws vary greatly from country to country, what's prohibited by one nation can be entirely permissible in another. Unless an international treaty applies (e.g., the United Nation's Contracts for the International Sale of Goods or CISG), countries are free to apply their own, quite different laws. The court that feels it has the greatest "connections", or the greatest interest in protecting its citizens, can and will take charge. It is very possible that the courts in two countries could reach two entirely different results–and this has happened numbers of times.

As one example, a French judge ordered U.S.-based portal Yahoo to block Web surfers in France from an auction where Nazi memorabilia were sold, including a fine of 100,000 francs ($13,700 U.S. dollars) for every day of noncompliance. Although Yahoo's offering sales of Nazi items was legally protected in the United States under the U.S. Constitution, it voluntarily banned the sale of these items in response. Arguing that the French court had no jurisdiction over it, however, Yahoo quickly countersued in a California Federal District Court to overturn that decision's effect in the United States.

In late 2001, the U.S. court ruled that Yahoo didn't have to comply with the French court's order (*Yahoo! v. La Ligue Contre le Racisme et L'Antisemitisme,* 2001). It held that a U.S. court may not enforce a foreign order that violates the U.S. Constitution by "chilling protected speech that occurs simultaneously within our borders."Thus, the U.S. court held that Yahoo didn't have to comply with all the laws in other countries that conflicted with those of the United States. Shortly afterwards, French civil rights groups appealed this decision to the Ninth Court of Appeals, and this case may very well journey all the way to the U.S. Supreme Court. The problem is, given other countries entering this fray, which court is right, when, and under or with what final authority?

Let's assume a Web site provides in its standard terms that U.S. law applies and all lawsuits must be brought at its U.S. headquarters. A German consumer feels that he's received a "lemon" and sues—of course—in Germany. Which law is Germany going to apply? You're right. German courts more than likely will apply German law in this case (by their own statutes). If the company sues in the United States, what law is that court going to apply? Probably U.S. law, because this country has a material interest in those proceedings. Because no international Supreme Court exists to adjudicate private disputes, there is no real way to settle this problem unless the parties later agree to those procedures. If the parties had negotiated the applicable law and forum before that dispute arose, then that agreement generally would control.

Fundamental differences among the various countries abound that affect basic principles, whether it's the United States and its First Amendment or EU countries with their basic consumer privacy protections. France mandates the use of the French language for numbers of documents in that country, while the EU and Japan have enacted strong antispam laws. Germany provides for a 2-week right of recession on online purchases, the U.S. to the contrary in this case, as well. One way to solve these questions is for countries to pass an International Jurisdiction treaty that binds the signatory states. Such legislation is in the works but will take years to finalize—and this state of affairs is another reason why alternative dispute resolution is growing in importance. For further treatment of international cyberlaw areas, see Julia Gladstone's chapter.

CYBERLAW DISPUTE RESOLUTION

One cyberlaw fact of life stands out: resolving disputes arising from the Internet's global reach through litigation is complex, expensive, and loaded with unclear results. In response, the Net community is actively pursuing alternative dispute resolution techniques (ADRs) such as mediation and arbitration—both offline and online. Given their low cost, confidentiality, limited negativity, and speed in resolving cyberdisputes, the use of ADRs is accelerating among users.

Web sites and online operators actively promote ADRs in their agreements and Terms of Use provisions. ADRs have been used to settle all types of Internet disputes, whether between Web partners, competing sites, domain name holders, ISPs and their subscribers, copyright holders and copiers, and many other Net matters. Credit card companies use an ADR form when they use "chargebacks" to end a customer's complaint with an online seller. As discussed before, ICANN has established a worldwide arbitration procedure to resolve domain name "cybersquatting" disputes. The U.S. Digital Copyright Act basically provides for an administrative procedure in resolving copyright disputes.

One of the striking ADR advances has been the rise of online cybermediators who work primarily online. For example, one party contacts the cybermediator about the problem and the parties' inability to solve it; the mediator then contacts the second party. If both parties agree to use a mediated approach and accept the ground rules, then the online mediation begins. Typically, each party e-mails the mediator with their position or an acceptable settlement amount. The mediator then intercedes, shuttling back and forth electronically to reach a settlement.

Although the experience has been that not having an actual physical presence between the mediator and parties (i.e., not experiencing body language and "real-time" emotions with words having a stronger unanticipated impact when made by e-mail) can be a drawback of online mediation, the 24/7 availability at any time, low cost, and no need to travel have proven to be advantageous.

Online resolution has particular advantages with lower monetary claims. In financial disputes, each party e-mails the amount at which they would settle their claim. In these "mediations," the agreed rules can provide for three rounds or more of settlement offers. Each party has also agreed to lower its demands by an agreed percentage—let's say 10%. By the third round, if the sides are close (let's say within 20%, or by some other formula), then that difference is halved and a deal struck. This is a brilliantly simple, mathematically oriented solution with special advantages for low-figure disputes.

The leading player in providing this "double-blind bid" procedure is Cybersettle.com, which was awarded a U.S. Patent for that process. Other ADR service providers in cyberspace are ClickNsettle (NAM), SquareTrade.com, InternetNeutral, American Arbitration Association (http://adr.org), and SettleOnline, to name a few. Rather than being caught up in establishing expensive legal precedents over simply the issue of which law applies, where, and when—and then the main legal case must be fought—more and more parties are settling their disputes on or off the Net by using ADRs.

THE LAW OF LINKING

The World Wide Web depends on linking for its existence, because this makes the Internet what it is. With the Net's maturity, however, the previous unconditional freedom to link has evolved into a framework of commonsense legal and netiquette rules that dictate limits on this freedom.

The general rule is that one doesn't need permission to link directly to another site, provided there is no commercial gain or some competitive informational advantage brought about by that linkage (even for a nonprofit institution). It is clear, moreover, that users should receive permission when they are "deep linking" or "framing," if only as a courtesy—and whether one should ask permission before any linking is a question of cyberethics, quite distinct from the law and any of its requirements. (See Jennifer Lagier's chapter on cyberethics for more on this area.)

Clearly, any stated or implied representation by linking that another's work is yours would be trademark infringement (e.g., using their logo), unfair competition and libel (e.g., saying something is yours when it's not), or a violation of the covenant of "good faith" that's implied in netiquette. Linking to illegal content by itself may also be illegal; in *Universal City Studios v. Reimerdes* (2000), the court enjoined the defendants from creating links from their court-prohibited site to a number of other "mirror" sites.

When links bypass homepages, connecting instead to a page deep within that site, additional considerations become present. Lawsuits have been filed and settled in the plaintiff's favor in which the plaintiff complained over "deep links" bypassing the advertising on their homepage, decreasing the "hit count" (users surf past the "count" page), diminishing their site's value, and allowing the defendant to "pass off" that information as its own.

The U.S. legal community, for example, watched closely when the owners of a newspaper, the *Shetland Times* in Scotland, brought a lawsuit against the *Shetland News,* a startup news service located in the same town. The *Shetland Times* published a daily online version of its newspaper, while the *News* was the first local daily to publish solely on the Web. It linked directly into the *Times* for news, and the *Shetland Times* went to court. The court granted the *Times* a temporary restraining order against the *News* and its linking practice (*Scottish Outer House,* 1996), and the case soon settled out of court.

One well-publicized framing case was the lawsuit brought by various media companies (CNN, Time Warner, *USA Today,* the *Washington Post,* etc.) against Total News for its framing strategy. The media argued that the use of those frames, whereby Total News showed news stories taken from the plaintiffs with only its advertising displayed, violated their copyright and trademark rights. Total News reached a settlement before trial, agreeing not to frame any content and paid to link to their sites in a separate window (*Washington Post Co., et al. v. Total News,* 1997).

An accelerating Net phenomenon has been the rise of linking agreements in which a linked site pays for the exposure. These situations occur in two ways: (a) the linkage is in reality an advertising contract or customer referral agreement (see Amazon.com's "Associate" program); (b) the linked site commercially profits or otherwise benefits from a deep link. In both cases, a written linking agreement is essential.

Commercially profiting by deep linking or sophisticated software without the other site's permission is another growing legal area. Known as "robots," "bots," "spiders," or "crawlers," these automated software systems steam past home pages deep into data banks, gathering information and transporting copies of whatever is desired back to the host site. If done frequently enough, these "hits" can create a near simultaneous look at whatever data is out there.

The largest Internet auction service, eBay, filed a lawsuit in late 1999 against Bidders Edge, one of several Net auction search services. It had been accessing eBay's site up to 125,000 times daily (as much as 1.53% of the total daily requests to eBay) in searching out what was going on specifically at eBay's auctions. EBay promptly sued after not being able to work out a license agreement with Bidders Edge to pay for this continuing access.

The judge granted an injunction, agreeing with eBay's contention that Bidder's Edge and its robots were trespassing on eBay's site by using and diminishing the resources of eBay's computer systems without permission (*eBay v. Bidder's Edge, 2000*). Bidder's Edge quickly appealed, but just before the appellate court issued its decision, the two companies agreed to settle their lawsuit. EBay reported that the settlement prohibited Bidder's Edge from sifting through its site for information and that Bidder's Edge agreed to pay an undisclosed sum of money.

Even search engines can be subject to linking conditions and exceptions. For example, in *Kelly v. Arriba Soft Corp.* (2002) a unanimous Ninth Circuit panel ruled that an Internet link to a full-sized copyrighted picture on a photographer's Web site did not qualify as "fair use" under U.S. copyright law and could be prohibited. In this case, Arriba's search engine, Ditto.com, displayed both full-sized and reduced-sized photographic images. The court found that the creation and use of the reduced-sized pictures, called "thumbnails," were fair use and could be continued. Internationally, a Danish court in *Danish Newspaper Publishers Association v. Newsbooster.com* (2002) ordered an online news site, Newsbooster.com, to remove links from its site to articles on the Web sites of various newspapers, on the ground that the links violated the EU Database Directory Act and bypassed the newspapers' homepages—following past precedent, both U.S. and international.

The directions of the law of linking are clear: (a) The general rule is that users do not need permission to link directly to the home page of another site, provided they don't disparage, misrepresent, or misappropriate; (b) given that these facts aren't present, framing and deep linking as opposed to linking will more likely constitute a violation; and (c) deep linking in commercial situations, as opposed to noncommercial or personel ones, are likely to be violations in which (1) direct competitors are involved, (2) there is an advantage being taken by that linkage, and (3) there is an element of "unfairness" or bad faith on the part of the linking party. Furthermore, if data is misappropriated, misused, or passed off by another as its own, even nonprofit or similar noncommercial sites may have valid causes of action.

CYBERCRIME

Cybercrime flourishes on the Internet, whether it is fraud, phony investments, pornography, rigged auctions, computer stalking, or prohibited gambling (see *http://www.cybercrime.gov* for more). The advantages of the Net for all users can quickly turn into disadvantages for law enforcement. The ease of entry and ability to disconnect from the Web allow criminals to appear and disappear within seconds with their ill-gotten gains. Arresting criminals is further complicated by the myriad jurisdictions that cybercriminals can cross so quickly, the protection of rogue nations, and differing state or national laws that can make extradition difficult. In turn, the authorities have had to add technology patrols to their arsenal of weapons, and users must be ever on the alert to not become victims.

Although nations add protective laws over time (e.g., the United States with its Access Device Fraud Act, 18 U.S.C. 1029 (1984); Computer Fraud and Abuse Act, 18 U.S.C. 1030 (1986); Trademark Counterfeit Act, 18 U.S.C. 2320 (1984); and the various ones previously mentioned), the question of jurisdiction and enforcement is always raised in this context. With criminal statutes, states and countries look at jurisdiction from the point of view of their laws and interest in protecting their citizens. For example, if gambling (or the Internet sale of wine) is illegal in State A but not State B, then a Web site in State B could be prosecuted by State A for its allowing the residents of State A to use that site. The reasoning is that every Web operator has the ability to be in compliance with State A's laws by simply refusing to allow A's residents to break their state's laws (i.e., by filtering out State A users).

Enforcement is always another question. Located in the Bahamas or other locations where activities such as gambling are legalized, just how do you enforce State A' s judgment penalizing another in a foreign state or country, not to mention the inherent conflict of laws question (i.e., the Yahoo Nazi memorabilia decisions)? Unless there's increased cooperation among the differing authorities and criminal justice treaties agreed to, the First Amendment legal considerations by themselves will be voluminous. When property rather than an individual's freedom is concerned, courts seem to have fewer problems determining rights, especially as to property that has already been seized. For example, *in U.S. v.* $*734,578.82 in U.S. Currency* (2002) the Third Circuit Court of Appeals affirmed a lower court's decision and upheld the forfeiture to the U.S. of funds seized from certain bank accounts. The funds were used in connection with an overseas gambling operation run out of England, where this type of gambling was legal. The court held that a New Jersey company—which had placed the funds in New Jersey bank accounts, then sent them to foreign institutions that in turn conducted the gambling activities—had violated U.S. law making it a crime to "conduct" an illegal gambling operation. The money under this decision could be lawfully seized.

The horrors of September 11 brought more aspects and another dimension of Internet crime to the forefront, given the ability of terrorists to communicate, raise money, and transfer assets over the Net. Among various legislative proposals, the enactment of the United States' International Money Laundering Abatement and Financial Anti-Terrorism Act of 2001, or the "USA Patriot Act," illustrates these new thrusts. Among other provisions, the USA Patriot Act requires key financial sectors to implement programs that prevent their services from being used to launder money or finance terrorism. Entire new industries, such as operators of credit card systems, money transfer companies, check cashiers, security firms, insurance companies, and even casinos—whether online or not—now are encompassed by strict regulations that once included only banks. This act also amended various provisions of the U.S. Code to allow broad interceptions of electronic communications and seizure of customer records. Other nations have or are enacting similar legislation. Their courts, including those of the United States, are currently being asked to rule on how legal these restraints are on individual privacy and constitutional rights. See Chuck Jaeger's chapter on cyberterrorism for more information.

CONCLUSION

Hold onto your hats, the Internet hurricane of change is still howling—but inside your office or study. Legally, as well as technologically, there are more vibrant areas of change coming. For example, professional associations these days face Web sites that give information out to potential patients, clients, and customers on a global basis.

From medicine and accounting to lawyering and filling drug prescriptions, state licensing boards are taking issue with this "practice without a required state license." This area will continue to be well litigated.

Another area involves the ability of the Internet to cut out the middleman. Because this medium allows consumers to contact suppliers directly, the old ways of conducting business are being seriously challenged in the courts. Travel agents are suing airlines, wine distributors are litigating with wineries that sell direct (in effect, suing their own customers or suppliers), and offline textbook distributors are suing online retailers, not to mention the ever-increasing numbers of other industries that are litigating these types of developments. Although the consumer has benefited, it's clear that the legal industry also has.

There's no question that the megasites and huge portals (such as AOL and Yahoo) dominate the Internet, and that the question of antitrust will rear its head even higher in the future. From AOL's acquisition of gigantic Time Warner to the Covisint cyber-venture between the world's six largest car manufacturers and their suppliers, the Web trend continues towards greater concentrations of power.

In addition to Covisint, various large virtual trading, exchange service Web sites abound: Altra Energy Technologies (utilities), ChemConnect (chemical and plastics), Converge (high-tech and electronics), Exostar (aerospace and defense), FuelQuest (independent oil and gas), PaperExchange (paper), and Transora (consumer packaged goods industry), among others. We will see over time more highly publicized Justice Department legal actions challenging these super offline and online combinations.

With the increase in cyberlaw actions over time, the rise and accepted use of cybercourts will also become a reality, along with more jurisdictions and courts converting to public accessible, electronic record keeping and filing. In this direction, Michigan became the first state to create a specific cybercourt (Mich. Pub. Acts, 2001). The new cybercourt under this legislation is expected to become operational in late 2003/2004. This court has concurrent jurisdiction over commercial litigation in disputes where the amount in controversy exceeds $25,000. All filings are made electronically, whereas all actions, depositions, and court appearances are to be by "electronic communications," such as streaming video, audio, and Internet conferencing; the intent is that there will be no paper transmitted or physical interface between judge, litigants, or witnesses. Appeals can be made either to a new cybercourt of appeals or through a normal appellate court. For the latest information and developments on Michigan's cybercourt, see http://www.michigancybercourt.net.

Regardless of the new Internet legal controversies that will rise up further in this new millennium, three realities exist: (a) The legal concepts already in place have proven to be quite adaptable to these challenges; (b) the concepts of fair play, common sense, and netiquette are filling in the gaps through court decisions and statutes; and (c) the use of ADRs on the Net will continue to grow over time because of the inappropriateness of litigation to solve the cyberdisputes among the citizens of the world.

The Internet has enhanced our lives and challenged our laws. The legal system is continuing to meet the challenge, but the world is never again going to be the same.

GLOSSARY

Anticybersquatting Consumer Protection Act (ACPA) The U.S. statute (15 U.S.C. 1125, 1999) that protects trademark or service-mark holders (including the names of famous people) from those who register a mark's domain name or its equivalent with a bad faith intent to profit from that act (e.g., cyberpirates). It allows the trademark or service-mark holder to sue for actual damages, statutory damages (when actual damages are difficult to prove), and/or force the domain name to be transferred back.

"Click" or "Clickwrap" contracts An agreement whereby a party agrees to the terms and conditions of an online agreement by clicking on a button reading "I agree," or some wording to that effect, to indicate the requisite mutual assent to those conditions and understanding.

Communications Decency Act Section 230 of this act (47 U.S.C. 223, 1996) provides that an online service provider is not to be treated as a publisher for purposes of liability for defamatory postings by third parties, nor liable for defamation in such cases.

Cyberlaw The emerging body of law that governs cyberspace transactions and disputes, otherwise known as the "Law of the Internet."

Cyberpirates Persons or entities who register a domain name that is the valid trademark or service mark of another, intending to sell that registered domain name back to the legitimate mark holder at a profit. This term is similar to "cybersquatters" who register domain names ahead of such interest but wait (or "squat") on those names until offers to buy back those names are received from others.

Dash-Dot Name A domain name registration that is off by one "dash dot" from another. For example, "www.microsoft.com" is registered as "www.microsoft-.com." ICANN regulations and case law have invalidated these registrations.

Defamation A false statement made by some person or entity about another, either orally or in writing, that is published to a third party and wrongfully harms the injured party's reputation.

Digital Millennium Copyright Act (DMCA) The 1998 act that amended U.S. copyright law and included (a) a section prohibiting circumvention of encryption or security protections on copyrighted software to violate its copyright (17 U.S.C. 1201–1204) and (b) a section on online service provider liability (17 U.S.C. 512). The online provider provisions set down an administrative proceeding that is used to resolve copyright disputes over third-party postings with online servers and establishes a "safe harbor" liability protection for those providers.

Fair Use The U.S. Copyright Act provides that the "fair use" of copyrighted works involving purposes such as criticism, comment, news reporting, teaching, scholarship, or research is not copyright infringement. Thus, some copying or copyright use is legally permissible in the United States that ordinarily would not be allowable in other countries.

Jurisdiction The power of a court or governmental agency to hear a case and decide the rights of the people or entities that appear before it. This jurisdiction can be *in personam* (determining the rights of people or entities, wherever they reside) or *in rem* (determining the ownership rights to property that is located within the court's territorial limits, regardless of where the disputing parties reside).

Internet Corporation for Assigned Names and Numbers (ICANN) The nonprofit organization that oversees a wide range of Internet functions (once the responsibility of the U.S. government) and now managed by an international board of directors. Among other functions, ICANN promulgates policy on the registration of domain names, accreditation of new registrars, and implementation of domain-name dispute resolution policies.

Long-Arm Statute U.S. state statutes authorizing a local court to assert personal jurisdiction over a nonresident defendant residing outside that state, given certain factual circumstances being present (for example, causing injury within that state by an act that takes place in part or wholly inside it).

Netiquette An informal, essentially noncodified doctrine of "Web manners," courtesy, and cyberethics aimed at creating a system establishing what is or isn't acceptable conduct on the Net, regardless of what the law provides.

No Electronic Theft Act (NET) The U.S. act (17 U.S.C. 506(a), 1997) that provides there is an illegal infringement when pirated copyrighted material over a 6-month period has a retail value of more than $1,000 (a misdemeanor) or more than $2,500 (a felony), even though there is no monetary gain or economic motivation with the infringer.

"Opt In" and "Opt Out" The two distinct privacy policies used by Internet firms and Web sites, which may or may not be codified. With "opt-out" provisions, the user must take the affirmative step to say "no" or refuse permission to a Web site's collection and transmission of financial and other sensitive consumer information. With "opt-in" policies, the Web site must take the steps to gain the positive approval of a user before it can collect, transmit, or sell such private information. Marketing firms prefer "opt-out" policies or laws, because these are less marketing restrictive and put the burden on the user, not the site.

Prima Facie Evidence Evidence presented that indicates a strong presumption the given fact or evidentiary assertion is factually true.

Service Mark A word, name, logo, mark, device, or some combination used by any person or entity to identify and distinguish services performed by it from those of another (e.g., eBay's name and logo).

Trademark A word, name, logo, mark, device, or some combination used by any person or entity to identify and distinguish its goods from those of another (e.g., McDonalds' golden arches or Nike's winged shoe).

Trademark Dilution When an entity uses a trademark (or service mark) of another in such a way as to cause that mark's value to diminish over a period of time. Although there is no immediate confusion as to whose product the mark applies, there is trademark infringement, given that the legitimate mark loses its value. For example, one William Nike starts up and operates "Nike's Ladies of the Night" striptease bar joints in cities where the legitimate Nike shoe and apparel manufacturer operates.

U.S. Copyright Office The U.S. agency that oversees the registration and regulation of copyrights (see www.loc.gov/copyright).

U.S. Patent and Trademark Office (PTO) The U.S. agency that oversees the registration and regulation of patents and trademarks/service marks (see www.uspto.gov for further information).

World Intellectual Property Organization (WIPO) The specialized United Nations agency and intergovernmental organization that is responsible for promulgating and administering major international intellectual property conventions.

CROSS REFERENCES

See *Copyright Law; Cybercrime and Cyberfraud; Cyberstalking; Cyberterrorism; Digital Identity; International Cyberlaw; Internet Censorship; Internet Etiquette (Netiquette); Legal, Social and Ethical Issues; Open Source Development and Licensing; Patent Law; Privacy Law; Taxation Issues; Trademark Law.*

REFERENCES

American Libraries Association v. U.S., 201 F. Supp. 2d 401 (2002).

Ashcroft v. American Civil Liberties Union, 535 U.S. 564 (2002).

Bally v. Faber, 29 F. Supp. 2d 1161 (1998).

Calder v. Jones, 465 U.S. 783 (1984).

Children's Internet Protection Act, 20 U.S.C. 9101 (2000).

Child Online Protection Act, Pub. L. No. 105-277, Title XIV (1998).

Children's Online Privacy Protection Act, 47 U.S.C. 231 (1998).

Crispi v. The Microsoft Network, L.L.C., 323 N.J. Super. 118 (N.J. App.Div., 1999).

Danish Newspaper Publishers Association v. Newsbooster.com, Bailiff's Court of Copenhagen, July 5, 2002.

Dow Jones v. Gutnick, 2002 HCA 56 (2002).

Electronic Signatures in Global and National Commerce Act, 15 U.S.C. 7001 (2000).

Ferrera, G., Lichtenstein, S., Reder, M., August, R., & Schiano, W. (2001). *Cyberlaw: Text and cases*. Cincinnati, OH: South-Western.

Financial Services Modernization Act, 15 U.S.C. 6801 (1999).

Ford Motor Co. v. Lane, E.D., Michigan, No. 99-74205 (1999).

DVD Copy Control Assoc. v. Bunner, 93 Cal. App. 4th 648 (6th Dist., 2001).

eBay v. Bidder's Edge, 100 F. Supp. 2d 1058 (2000).

Eldredge v. Ashcroft, U.S. Sup. Ct. No. 01-618 (2003).

Geoff v. AOL, Inc., No. PC 97-0331, R.I. Super. Ct. (1998).

Graham v. Oppenheimer, No. 3:00cv57, E.D. Va. (2000).

Greenberg v. National Geographic Society, 244 F.3d 1267 (2001).

Health Insurance Portability and Accountability Act, Pub. L. No. 104-191 (1996).

International Money Laundering Abatement and Financial Anti-Terrorism Act, U.S. Pub. Laws No. 107-56, Title III (2001).

International Shoe Company v. Washington State, 66 U.S. Sup. Ct. 154 (1945).

Kelly v. Arriba Soft Corp., 280 F.3d 934 (2002).

Mattel v. Barbie-club.com, 310 F.3d 293 (2002).

Mich. Pub. Acts 262 (2001).

Multnomah County Public Library v. U.S., 201 F. Supp. 2d 401 (2002).

National Bellas Hess v. State of Illinois, 386 U.S. 753 (1976).

New York Times v. Tasini, 533 U.S. 483 (2001).

Parisi v. Netlearning, 139 F. Supp. 2nd 745 (2001).

Planned Parenthood v. American Coalition of Life Activists, 290 F.3d 1058 (2002).

Pavlovich v. Superior Court (DVD Copy Control Association, real party in interest), 29 Cal. 4th 262 (2002).

Recording Industry Association of America v. Verizon Internet Services, 2003 U.S. Dist. LEXIS 681 (D.D.C., 2003).

Register.Com, Inc. v. Verio, 126 F. Supp. 2nd 238 (2000).

Reno v. American Civil Liberties Union, 521 U.S. 844 (1997).

Sallen v. Corinthians Licenciamentos Ltda., 273 F.3d 14 (2001),

Shetland Times v. Shetland News, Scottish Outer House, 1997 S.C. 604 (1996).

Sonny Bono Copyright Term Extension Act, 17 U.S.C. 101, 302–305 (Supp. 1999).

Specht v. Netscape Communications Corp., 306 F.3d 17 (2002).

State Street Bank v. Signature Financial Group 149 F.3d 1368 (1998).

Universal City Studios v. Reimerdes, 82 F. Supp. 2nd 211 (2000).

U.S. v. $734,578.82 in U.S. Currency, 286 F.3d 641 (2002).

Virtuality v. Bata, 138 F. Supp. 2nd 677 (2001).

Washington Post Co., et al. v. Total News, 97 Civ. 1190, S.D.N.Y. (1997).

Weber-Stephen v. Armitage, 2000 U.S. Dist. LEXIS 6335 (2000).

WIPO Copyright Treaty, 36 I.L.M. WIPO Treaty, WIPO Doc. CRNR/DC/94 (1996).

Yahoo! v. La Ligue Contre le Racisme et L'Antisemitisme,169 F. Supp. 2nd 1181 (2001).

Zippo Manufacturing Co. v. Zippo Dot Com, Inc., 952 F. Supp. 1122 (1997).

FURTHER READING

Consumer Sentinel, for the latest on primarily U.S. fraud protection and online complaint procedures: http://www.consumer.gov/sentinel.

Department of Commerce, for latest developments including EU "safe harbor" privacy guidelines: http://www.doc.gov.

Econsumer.gov, for data on international fraud protection and online complaint procedures: http://econsumer.gov.

For EU developments, including its copyright directive: http://eurorights.org.

FTC site, for data on COPA, consumer protection, and latest developments in other areas: http://www.ftc.gov.

ICANN, for information on domain name dispute resolution process, registration, and accepted registrars: http://icann.org.

Powers, Dennis M. (2002). *The internet legal guide: Everything you need to know when doing business online*. New York: Wiley.

Secruities and Exchange Commission, for information on securities and investment fraud: http://www.sec.gov.

Spamlaws.com, for the latest developments on anti-spam laws, both in the United States and globally: http://spamlaws.com.

U.S. Copyright Office information, including DCMA and registration data: http://www. loc.gov/copyright/.

U.S. Dept. of Justice, for overall and latest information on computer-related crimes: http://www.cybercrime.gov.

U.S. Patent and Trademark Office, for information on trademarks, service marks, and patents: http://www.uspto.gov.

Cyberterrorism

Charles W. Jaeger, *Southerrn Oregon University*

WHAT IS CYBERTERRORISM?

Incidents in Cyberspace

In November 1999, a Tampa, Florida man was charged in Federal court with using the Internet to threaten to destroy the reputations of six young women and girls unless they engaged in phone and cybersex. The incident was investigated by a task force that included the FBI (FBI, 1999).

In March 2001, a Canadian known as "Mafia Boy" answered charges related to large-scale "denial of service" (DOS) attacks causing $1.7 billion damage in which millions of e-mail messages brought down some of the Internet's largest e-commerce sites, including Amazon, Yahoo! and eBay. He pled guilty to 56 charges but boasted that he will commit other such acts in the future (Raines, 2001). Mafia Boy was a juvenile.

Perhaps the world's most notorious cybercriminal and the first to appear on the FBI's "Most Wanted" list, "the Condor" stole tens of thousands of credit card numbers and copied millions of dollars worth of computer software beginning in the 1980s. Kevin Mitnick eventually served five years in prison (Meriwether, 1995; Sargent, 2001).

The Terrorism Research Center's Information Warfare Database lists over 50 "incidents" dating back to 1982 of unauthorized entry, denial of service attacks, and similar cyberspace events. Targets include the North American Air Defense Command, NASA, U.S. military sites, the White House, the U.S. Department of Defense, and many others vital to U.S. and world security (IWDB, 2002).

People are becoming concerned that criminal behavior in cyberspace can affect them directly. A December 2001 survey reported that 74% of Americans expressed fear that their personal information on the Internet could be stolen or used for malicious purposes and are concerned that a cyberattack could target critical infrastructure assets like telephone networks or power plants. Harris Miller, President of the Information Technology Association of America (ITAA), said, "The attacks of Sept. 11 . . . destroyed peace of mind for many people using the Internet [and] is generating high anxiety in cyberspace" (Greenspan, 2002a).

While these are mostly "hacking" or "cybercrime" incidents, people increasingly recognize the potential for large-scale attacks that have the potential to create terror, and these prospects evoke fear. Yet, of the above cases, only the cybersex case resulted in charges of cyberterrorism. That incident included no unauthorized break-ins, affected relatively few people, and created virtually no property damage or other direct monetary losses. However, the FBI's Web site calls the attacks "large scale," and Frank Gallagher, Special Agent in Charge of the Tampa Division of the FBI, described them as "cyberterrorism" because of the fear created in the recipients.

Problems Defining Cyberterrorism

Acts such as these may or may not constitute cyberterrorism. It is important to differentiate acts of cyberterrorism from vandalism, common hacking, and criminal activity. Cyberspace is so new and rapidly changing that it is difficult to be precise. An overly broad definition will include so many acts that the term "terrorism" would become meaningless. Making a definition too detailed, in contrast, risks omitting acts that should qualify.

Fifty years ago, there was no mention of "cyber" in *Webster's Dictionary.* Today, it is used as a prefix for topics related to computers and/or networks, including acts of terror (http://www.dictionary.com). The term "cyberterrorism" is coming into common usage, and an evolving body of knowledge is moving toward defining it. Separate government departments and agencies have used this term, but none have formulated a definition of cyberterrorism that is binding outside their sphere of influence.

Definitions in common usage are seldom precise enough to be useful. A 1998 student paper written for a computer ethics class defined cyberterrorism as "the use of computing resources to intimidate or coerce others" (Sproles & Byars, 1998). At a conference early in 2002, Howard Schmidt, vice-chairman of the new Critical Infrastructure Protection Board and formerly chief security officer for Microsoft Corp., defined cyberterrorism as "anything that disrupts or causes mistrust about the security of computers and networks" (Moran, 2002). These two broad interpretations would include many forms of cybervandalism, hacking, and the like. While these activities might be annoying or damaging, they are probably not cyberterrorism.

The National Infrastructure Protection Center (NIPC) is a group of over 100 special agents, analysts, and others from the FBI, the Department of Defense, the CIA, the National Security Agency, and other federal departments (Vatis, 2001). It calls cyberterrorism an "evolving concept" and states, "the definition of terrorism must evolve to reflect the type of activity that goes beyond traditional physical violence" (NIPC, 2001b). The NIPC's Analysis and Information Sharing Unit has proposed the following definition:

> Cyber-terrorism is a criminal act perpetrated by *the use of* computers and telecommunications capabilities resulting in violence, destruction and/or disruption of services to create fear by causing confusion and uncertainty within a given population, with the goal of influencing a government or population to conform to a particular political, social, or ideological agenda. (NIPC, 2001b; emphasis added)

In congressional testimony, an official of the FBI's Counterterrorism and Counterintelligence unit defined cyberterrorism as "*the use of* cyber tools to shut down critical national infrastructures . . . for the purpose of coercing or intimidating a government or civilian population" (Greenspan, 2002b; emphasis added).

Mark Pollitt, of the FBI Laboratory in Washington, DC, in connection with his private academic studies at George Washington University, constructs a "working definition" as follows:

> Cyber-terrorism is the premeditated, politically motivated attack *against* information, computer systems, computer programs, and data which result in violence against noncombatant targets by sub national groups or clandestine agents. (Pollitt, 2002; emphasis added)

Dorothy Denning, Director of the Institute for Information Assurance, Georgetown University, expands the definitions by stating, "The attack should be sufficiently destructive or disruptive to generate fear comparable to that from physical acts of terrorism. Attacks that lead to death or bodily injury, extended power outages, plane crashes, water contamination, or major economic losses would be examples" (Denning, 2001).

Each of the definitions addresses a type of terrorism that goes beyond traditional physical violence and into cyberspace, but there are important differences. The NIPC emphasizes *the use of* computers in creating terror, emphasizing the act, itself. Pollitt emphasizes the *target* of the terror, or the victim(s). A logical definition of cyberterrorism would encompass three elements: the *use*, the *victim*, and the concept of *terror* itself in acts that may or may not include physical violence.

DIFFERENCES BETWEEN CYBERTERRORISM AND OTHER TERRORISM

Terrorism Definitions

Terrorism in the physical world has existed throughout history in one form or another. Since there is no consistent definition of cyberterrorism, it is reasonable to consider one by beginning with physical world acts of *terror*. Again, however, there is some inconsistency. The FBI and the CIA each have their own definitions of "terrorism."

The CIA's Counterterrorist Center refers to Title 22 of the U.S. Code, Section 2656f(d), in defining terrorism as "premeditated, politically motivated violence perpetrated against noncombatant targets by subnational groups or clandestine agents, usually intended to influence an audience" (DCI Counterterrorist Center, CIA, 2002).

The FBI defines domestic terrorism as "the unlawful use, or threatened use, of violence by a group or individual . . . to intimidate or coerce a government, the civilian population, or any segment thereof, in furtherance of political or social objectives" (Watson, 2002).

The two versions share common references to politics or politically motivated violence, but there are differences. One specifies actual *acts*, while the other could include *threats*. One mentions premeditation. The other specifies subnational groups or clandestine agents. These differences may seem trivial, but they become important when applied to specific cases of guilt or innocence.

Each of the acts cited at the beginning of this chapter concern some variation of fear, confusion, uncertainty, violence, intimidation, or coercion, and each at least implies a form of intent. None, however, including the one actually charged with cyberterrorism, is an ideal fit with the above definitions. The acts all lack overt political motives, none resulted in physical violence, and in most cases, there is a lack of intense fear, dread, and panic.

The Role of Fear, Panic, Violence, and Intent

Fifty years ago, *Webster's Dictionary* defined terrorism as the "use of terror and violence to intimidate, subjugate, etc., especially as a political weapon or policy." Terror was defined as "intense fear" or "the quality of causing dread;

terribleness." An example specified "political executions, as during revolution" (Webster, 1952). Recent versions of *Webster's* have changed very little. However, neither of the modern definitions used by our security agencies includes the elements of intense fear, dread, or panic, although they may be implied through intimidation. Both agencies specify violence or threatened violence.

All of the definitions of terrorism and cyberterrorism imply some variation of intent. These are not random or accidental acts and threats. The CIA definition specifies "premeditated" acts, while the FBI definition implies intention. Terrorism is usually considered to include some form of collective intent by a group or organization with an ideology intended to cause harm (FBI, 1999).

The Worldwide Perspective

Destructive acts can occur anywhere in the world. Indeed, terrorism and cyberterrorism have become worldwide concerns. Following the September 11, 2001, attack on the World Trade Center, the 56th regular session of the United Nations General Assembly unanimously passed a resolution condemning terrorism. More narrowly, the resolution includes cyberterrorism. One purpose of the UN is to maintain international peace and security. Clearly, terrorism threatens UN goals, and they recognize that cyberspace can be a weapon against them.

"The same Internet that has facilitated the spread of human rights and good governance norms has also been a conduit for propagating intolerance and has diffused information necessary for building weapons of terror" (Annan, 1999).

Although there are differences between the way individual nations and the world community view cyberterrorism, there is general agreement that it has the potential to create chaos and contribute to a lower standard of living on a global scale.

International agencies and other groups are beginning to work toward neutralizing cyberterror weapons. On November 23, 2001, the first-ever international treaty on criminal offences committed against or with the help of computer networks was signed in Budapest, Hungary, by 26 member states of the Convention on Cybercrime, a part of the Council of Europe. Four nonmembers who helped draft the document, Canada, Japan, South Africa, and the United States, also signed the treaty. The Convention deals with various cybercrimes and network security. Its main aim is to pursue "a common criminal policy aimed at the protection of society against cyber-crime" (Convention on Cybercrime, 2001a). Christian Kruger, Deputy Secretary General of the Council of Europe, announced that "the Convention would give national legal systems ways of reacting together to crimes committed against or through computer networks, especially those related to terrorism" (Convention on Cybercrime, 2001b).

WHO DECIDES WHETHER AN ACT IS TERRORISM?

Although there is no universal definition of terrorism or cyberterrorism, the terms are a basis for laws and regulations. In the United States, crimes and damages are usually defined beginning with legislation. To implement legislation in detail, government agencies enact specific regulations, and the regulations often define specific terms. Since cyberlaw is in its early stages, interpretation becomes more difficult.

Individuals, businesses, and other parties that disagree with the law or the agency implementations may challenge them in federal, state, or local courts. Courts consider definitions at every level in rendering their judgments. A trial court makes an interpretation based on the hermeneutics of the law and prior case law that has concerned the same or similar issues. Dissatisfied parties may appeal the ruling to a higher court for review. Ultimately, a federal or state supreme (or other highest) court makes a determination for a given situation that is binding on all parties and becomes a basis for future interpretations. Changes require new legislation or regulations, and the process begins anew. In the Florida cybersex case cited at the beginning of this chapter, the courts may ultimately determine whether cyberterrorism or simple criminal activity was involved.

Legislators, regulatory agencies, and the courts are all handicapped when interpreting potential cyberterrorism. Traditionally, laws have dealt with acts that occur in time and space. Crimes in cyberspace may have physical effects in time and place, but some of their causal factors reside in a virtual reality, mostly independent of a particular time and space. Laws and regulations are typically written by people who are more comfortable referencing time and space, and there is a scarcity of case law applying directly to causal factors residing outside of these parameters. Many other nations have similar systems—and interpretation problems.

This chapter cannot fully explore a precise definition of cyberterrorism. However, using terminology from the various agencies, it is reasonable to think that evolving cyberterrorism definitions will include acts designed to create fear, dread, or panic that may be directed at noncombatant targets by subnational groups or clandestine agents to intimidate or coerce a government or some segment of the civilian population in furtherance of political or social objectives. This chapter proceeds from this somewhat imprecise point.

HOW AND WHY WAS CYBERTERRORISM CREATED?

The Relationship to Cybervandalism, Hackers, and Other Cybercrimes

Using a broad definition, many ordinary crimes, vandalism, and hacking might be classified as terrorism. Even repeated "spam" might be cyberterrorism if it intimidates or coerces people into purchasing or otherwise acting on an offer. Pollitt's definition narrows the target to politically motivated attacks against digital devices. That would exclude threats against individuals or businesses such as the cybersex and Mafia Boy cases cited above, incidents that had a substantial fear or even panic component, but were not overtly political. The NIPC definition narrows it further, but it also excludes the cybersex case, the only case where actual charges of terrorism were brought, because the perpetrator did not have an overt political, social, or ideological agenda. The cyberterrorism term is widely misused.

Who Are Today's Cyberterrorists?

President Clinton, in a December 2000 address at the University of Nebraska, said, "One of the biggest threats to the future is going to be cyberterrorism—people fooling with your computer networks, trying to shut down your phones, erase bank records, mess up airline schedules, do things to interrupt the fabric of life" (Clinton, 2000). These kinds of incidents and many other activities in cyberspace are annoying and cause monetary or other damages. They include some degree of vandalism, mischief, unauthorized entry, or other destructive act(s), but only in a broad sense could they be construed to create fear, intimidation, or panic. There is considerable "gray area" when we consider actual cases.

In July 2002, CNN reported that Yale University complained to the FBI that it experienced 18 unauthorized log-ins to their Web site that were traced back to computers at Princeton, including computers in the admissions office (CNN.com, 2002c). In 1999, after the United States accidentally bombed the Chinese embassy in Belgrade, Yugoslavia, U.S. Web sites including the Department of Energy, the Department of the Interior, and the National Park Service were defaced in the name of China, and the White House Web site was shut down for three days after their servers were overloaded with massive amounts of e-mail. In August and September of that year, pro-Chinese hackers defaced or otherwise compromised 165 Taiwanese Web sites to protest the Taiwanese presidential elections (Jane's). In May 2001 some of the same activists attacked U.S. Web sites after a Chinese fighter jet was lost following a collision with a U.S. reconnaissance plane (Tang, 2001). In 2000, pro-Pakistani hackers defaced more than 500 Indian Web sites. In April 2001, pro-Korean hackers, primarily university students, attacked the computers and Web sites of various Japanese organizations following their approval of a new history textbook that the Koreans believed excluded atrocities committed by Japan in connection with World War II (NIPC, 2001a).

Each of these incidents created trouble or made a symbolic statement, and these kinds of people are widely referred to as cyberterrorists. Some are politically motivated, but they don't really create widespread fear, dread, or panic. While the target might be similar, and the same techniques can be used, this is not cyberterrorism. The objective, degree of violence, and reaction are different.

Conway describes a useful three-tiered schema for categorizing "fringe" activity on the Internet as "use, misuse, and offensive use" (Conway, 2002). Use is normal and legal Internet usage—the same for terrorists as for anyone else. Misuse includes acts that disrupt or otherwise compromise other sites, including vandalism and protests typically associated with hackers. Most of the above incidents fall into this category. Offensive use entails actual damage, theft, fraud, extortion, or commercial espionage, much of which is criminal. A subset of offensive use would be cybercrime, and a subset of cybercrime would be cyberterrorism. See Figure 1.

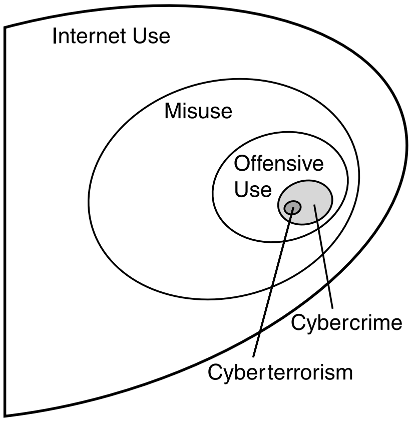


Figure 1: Categories of Internet usage.

Many misuse and offensive use "incidents" are tracked by the Carnegie Mellon CERT Coordination Center. The incidents counts have increased exponentially since 1988 (Carnegie Mellon CERT Coordination Center, 2002). See Table 1 and Figure 2.

Terrorists seldom engage in ordinary misuse or offensive use, as it is counterproductive to their larger goals through the potential risk of discovery. They engage in a small but potent subset of cybercrime. Today, the cyberterrorist threat that concerns the FBI, CIA, UN, and almost everyone else entails an end result including widespread death, destruction, fear and dread, chaos, or any of several other devastating consequences. Because of the magnitude of the potential effects, this small subset of Internet offensive use is a new and larger threat than the world has experienced.

Acts That May Be Cyberterrorism

Barry C. Collin, who takes credit for coining the term "cyberterrorism" in the mid-1980s, says, "This enemy does not attack us with truckloads of explosives, nor with briefcases of Sarin gas, nor with dynamite strapped to the bodies of fanatics. This enemy attacks us with ones and zeros" (Collin, 1996). The outcome can be deadly. He lists some possible acts of the cyberterrorist:

Table 1 Number of Incidents

Year	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002*
Incidents	6	132	252	406	773	1,334	2,340	2,412	2,573	2,134	3,734	9,859	21,756	52,658	102,706

*2002 estimated based on 73,359 incidents reported through Q3.
Note: Data from http://www.cert.org/stats/.

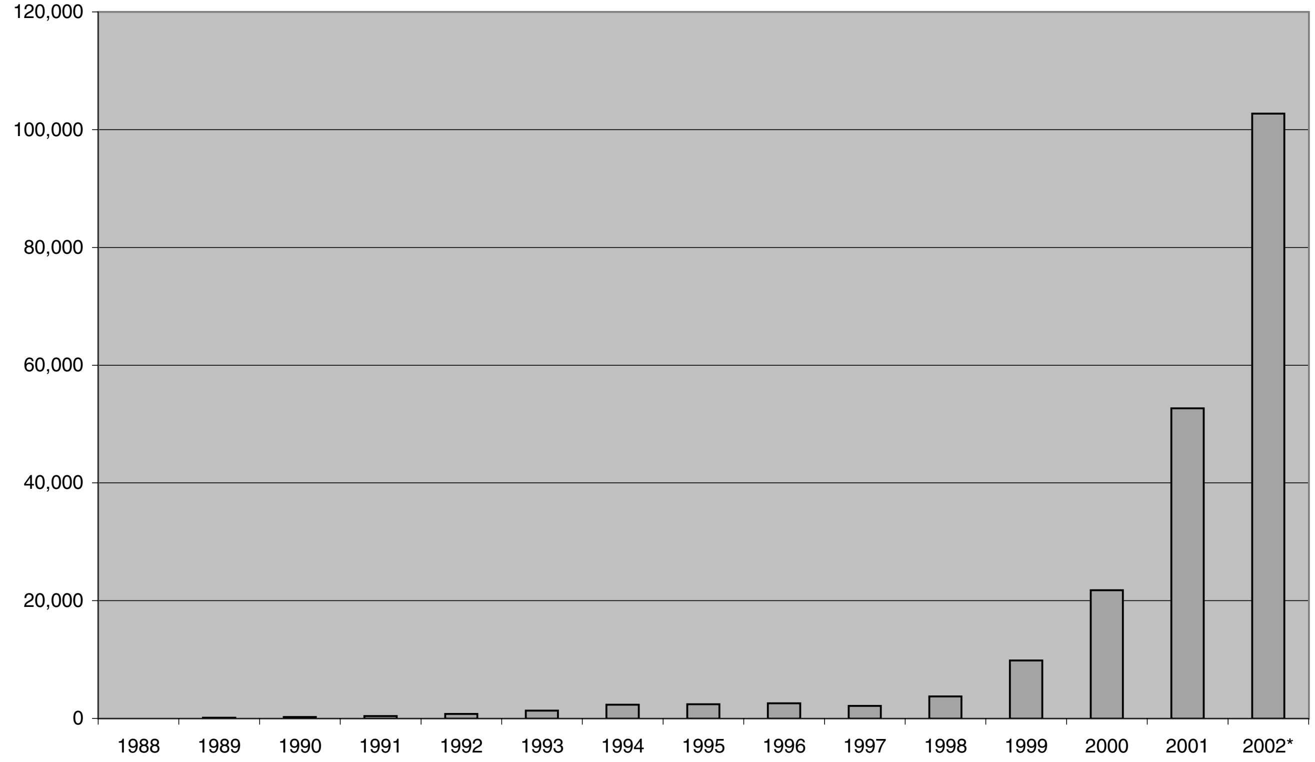


Figure 2: Number of incidents (graphical representation of Table 1).

Remotely changing the pressure in gas lines, causing valve failures, explosion, and fire.

Placing computerized bombs around a city, all simultaneously transmitting unique numeric patterns, each bomb receiving each other's pattern. If any of the set of bombs stops transmitting, all the others detonate simultaneously, which effectively prevents disarming any of the bombs.

Attacking future air traffic control systems to cause civilian jets to collide.

These "ones and zeros" attacks would have violent and fearful physical effects equally or more devastating than those caused by the truckloads of explosives, dynamite, or poison. Following the 9/11 attacks on the World Trade Center, it was reported that the government asked some of the film industry's most creative minds to come up with additional scenarios for potential attacks so that safeguards could be considered. Cyberterrorists might open a dam's spillway to inundate downstream communities or cause a meltdown of a nuclear power plant.

Collin also lists several acts in cyberspace that do *not* have devastating physical effects, but clearly have the potential to create widespread chaos, dread, or intimidation in offices, homes, and on the streets:

Remotely accessing the processing control systems of a cereal manufacturer to alter the formula and sicken children;

Disrupting banks and international financial institutions and stock exchanges, with resulting loss of confidence in the economic system;

Remotely altering formulas of medication at pharmaceutical manufacturers, resulting in ineffective or potentially harmful medications; and

Shutting down the electrical grid, causing wide-spread chaos.

Collin concludes, "In effect, the cyber-terrorist will make certain that the population of a nation will not be able to eat, to drink, to move, or to live. In addition, the people charged with the protection of their nation will not have warning, and will not be able to shut down the terrorist, since that cyber-terrorist is most likely on the other side of the world."

In November 2001, Richard Clarke, chairman of President Bush's Critical Infrastructure Protection Board, said cyberattacks on the nation's critical [information technology] infrastructure could potentially cause "catastrophic damage to the economy" akin to the "functional equivalent of 767's crashing into buildings" (Johnson & Radcliff, 2001).

What may make these acts even more dreadful is that most of them could be carried out from remote locations, and the perpetrators would be difficult to track down. "This is the new world of cyber-terrorism," says Tom Regan in an article for the *Christian Science Monitor*. "Cyber-terrorism allows terrorists—both foreign and domestic—to inflict damage with no harm to themselves and little chance of being caught" (Regan, 1999).

The Computer Emergency Response Team (CERT) Coordination Center, part of the Software Engineering Institute, is operated by Carnegie Mellon University for the

Department of Defense. Dr. Howard Lipson, in a November 2002 special report, says:

> The current state of the practice regarding the technical ability to track and trace Internet-based attacks is primitive at best. . . . Society continues to migrate increasingly critical applications and infrastructures onto the Internet, despite severe shortcomings in computer and network security and serious deficiencies in the design of the Internet itself. Internet protocols were designed for an environment of trustworthy academic and government users. . . . In this era of open, highly distributed complex systems, vulnerabilities abound [and] existing track and trace capabilities are primitive compared with the capabilities of attackers. (Lipson, 2002)

In 1998, Michael Vatis, director of the NIPC, was reported to have told a Senate subcommittee, "Tracing cyber-attacks is like 'tracking vapor' " (Christensen, 1999). Unknown and unseen, an enemy difficult to bring to justice connotes a lack of control that may make people more fearful.

Why Do They Do It?

Fun and Games, Attention and Peer Group Recognition, Profit

Some of the acts described by President Clinton and the Yale versus Princeton incidents may be nothing more than leisure activities by individuals that find them fun. Some of the incidents involve simple mischief, rather than offensive use. Many hackers engage in misuse activities to be recognized by their peers. Several accomplished hackers have been hired by business or government entities to prevent hacking, enhance network security, or edit magazines or newsletters that relate to these topics.

Revenge and Political Statements

In October 2001, the NIPC issued a report on rapidly growing worldwide activities designed to "expose government corruption or fundamental violation of human rights" or achieve similar goals. These include long-term attacks characterized as "hacktivism." "Cyber protests" are short periods of intense political activity. Both are used for revenge:

> In the last decade, with the explosion of the size of the Internet, protests and political activism have entered a new realm. . . . Cyber protests have become a worldwide phenomenon. . . . Unrestrained by geographic boundaries, protesters have an enormous forum in which to be heard. . . . It has only been since 1998 that cyber protests have skyrocketed in popularity and become commonplace in today's computerized world. (NIPC, 2001a)

Much as a delinquent child might act out to receive attention from a parent, teacher, or peer group, hackers can create havoc designed to make the world take notice of their opinions and belief. Defacing Web sites and interrupting e-commerce can highlight causes and attract the media spotlight, while it creates embarrassment and monetary costs for their adversaries.

Destruction, Fear, Dread, and Panic

Cyberterrorists usually have a political or societal agenda that is much more comprehensive and severe than a cyberprotester's. They may feel intense hatred for a business, society, government, way of life, or certain religions. They may seek to punish these targets—including individuals and entire populations not directly involved in the policy decisions with which they disagree. These acts almost always involve offensive use.

Terrorists typically believe strongly in their cause. They may attract a small following, and their methods may be condemned, or at least not widely supported. In this sense, they are extremists or fanatics. To communicate their point of view, they may create chaos through some horrific act, leaving fear, dread, and panic. In the vacuum that follows, they may hope to introduce their desired changes.

The Role of Asymmetric Response

Asymmetric response is a generic explanation. In free countries, those who do not approve of a government, business, or institution policy can disagree publicly. If their position is widely unpopular or repugnant, they would be considered "extreme." They may become frustrated, feel powerless, and believe that it is "them against the world." Convinced that they are right and unable to achieve their objectives through conventional means, they can strike against the "enemy" with an act of terror. This is the asymmetric (one-sided) response.

Asymmetric responses often are associated with conventional warfare. A rag-tag, poorly trained militia or even a neighborhood gang that is no match for a powerful state army may resort to guerilla tactics, sneak attacks, bombings, sniper attacks, or even rock throwing to harass and demoralize the superior force. This can be highly effective, as witnessed by Russia's withdrawal from Afghanistan and the United States' departure from Somalia.

The National Intelligence Council (NIC) has predicted asymmetric responses to an expected overwhelming military superiority brought about by a paradigm shift in the nature and conduct of warfare. This "revolution in military affairs" (RMA) will be "a small, information-intensive, professional armed force [as] the model for a 21st century military," largely based on strength in information technology and smart weaponry. As the balance tilts more and more toward the camp of the powerful adversary, asymmetric response becomes more likely, even between nations with broadly similar tools. "Iraq and Iran are examples of states that will likely explore the usefulness of information technology in pursuit of asymmetric conflict . . . including through the employment of information warfare and cyber-terrorism" (NIC, 1999).

Military analyst Tony Cordesman, former adjunct professor of national security studies at Georgetown

University and author of numerous books and articles dealing with asymmetric response predicts, "It is very likely that over the next few years, the [U.S. National] Guard will greatly expand its mission . . . [and] may well be reconfigured to perform more missions in asymmetric warfare" (ABC News, 2001).

Asymmetric cyberterrorism may be next. The Associated Press reported that White House Technology Advisor Richard Clarke, testifying before a Senate Judiciary subcommittee on cyberterrorism on February 13, 2002, said, "A serious cyber-attack is almost inevitable because it is cheaper and easier for a foreign country or a terrorist group than a physical attack" (Holland, 2002).

Denning notes that the Internet "may lessen the need for violence by making it easier for sub-state groups to get their message out," but concludes, "Unfortunately, this . . . does not seem to be supported by recent events. . . . Groups that foster hate and aggression thrive on the Internet alongside those that promote tolerance and peace" (Denning, 2001).

The industrialized nations have highly developed electronic infrastructure that drives knowledge-based economies and gives them advantages in business, government, and military power. The virtual world of cyberspace increasingly leaves others behind, including some who consider technological progress in fundamental opposition to their belief systems or religions. More and more powerless, more extreme in their beliefs, they can neutralize their enemies' strength through asymmetric cyberattacks. Regan states, "Seemingly unconnected events may have a more sinister source: coordinated cyber-hacker attacks." He adds, "No other country or group can approach the US conventional-weapon superiority. This is why many terrorists find information terrorism an attractive alternative to traditional forms of terrorism . . . . It is a way for the 'weak' to attack the 'strong' " (Regan, 1999). The contradiction is that they may be using the very tools they condemn.

WHO SPONSORS CYBERTERRORISM?

When the world was simpler and more predictable, it was easier to determine who was pulling strings and causing consequences. Today's terrorists can pursue extreme agendas with more powerful tools than ever before. Many of these—especially those in cyberspace—are amazingly inexpensive, easy to get, and easily concealed.

In simpler times, terrorists required a sponsoring organization—often state sponsorship—to obtain sufficient funding and connections. Today, cyberterrorists acting alone or in small groups can pursue extreme agendas either through volunteer labor or by relatively small amounts of money disbursed from wealthy individuals, foundations, fund raising organizations, or almost any rogue country's intelligence unit. "Official" state sponsorship is less likely, and following a money trail is elusive in that physical currency need not be exchanged. Regan quotes Dr. Harvey Kushner, chairman of the criminal-justice department at Long Island University:

> We have moved away from state-sponsored terrorism. The old model of the hierarchical or "organized crime" group, no longer exists. These days, terrorists move in loose groups, constellations with free-flowing structures. So these days terrorism—both the traditional kind and cyberterrorism—is more the act of the freelancer or the individual. This is true both internationally and nationally. (Regan, 1999)

WHO ARE MOST VULNERABLE TO CYBERTERRORISM?

Cyberterrorism Vulnerabilities Today

It has been reported that "The U.S. Department of Defense logged 250,000 computer attacks on its sites in 1996 alone, 62 percent of them successful" (Thom, 1999).

In 1997, 35 hackers were hired by the National Security Agency to launch operation "Eligible Receiver," an exercise of simulated attacks on the U.S electronic infrastructure. They reportedly achieved "root level" access in 36 of the Department of Defense's 40,000 networks, "turned off" sections of the U.S. power grid, "shut down" parts of the 911 network in Washington, DC, and other cities, and broke into computer systems aboard a Navy cruiser at sea. Richard A. Clarke, White House terrorism czar, called this an "electronic Pearl Harbor" (Christensen, 1999).

On May 4, 2001, the White House Web site was the victim of a distributed denial of service (DDOS) attack lasting about three hours that effectively shut it down by overloading its servers with an automated barrage of service requests. This was symptomatic of similar attacks on commercial and other government sites, some of which may have included the Chinese hackers attacks (Weiss, 2001).

In late August, 2002, CNN published an Associated Press report about an FBI raid on the offices of a consulting firm in San Diego, CA, following a local newspaper story about the company's claim that it had found security loopholes in U.S. military computers, some of the country's most confidential data repositories. The company "identified 34 military sites where it said network security was easily compromised [using] 'free' software and relatively standard hacking procedures to identify vulnerable computers and then peruse hundreds of confidential files containing military procedures, e-mail, Social Security numbers and financial data." The company claimed they were simply trying to expose a need for better security. The FBI and Justice Department were not amused (CNN.com, 2002d).

On October 21, 2002, an estimated 6,000 computers worldwide swamped the Internet's 13 root servers that control most of the Net's traffic in a distributed denial of service attack. Eight were disabled to some degree in what was called "the biggest ever hacking attack on the Internet" (Naraine, 2002; Vickers, 2002).

On December 5, 2002, federal agents raided Ptech, a Quincy, MA, company "that provides critical software to major U.S. agencies and is suspected of having ties to Usama bin Laden and Al Qaeda terrorists . . . . Officials had suspected 'back doors' may have been built into Ptech software that could enable terrorists to access federal computers." Ptech's customers include the Department of Energy, the FBI, the U.S. Air Force, the Naval Air Systems

Table 2 Number of Vulnerabilities

Year	1995	1996	1997	1998	1999	2000	2001	2002*
Vulnerabilities	171	345	311	262	417	1,090	2,437	4,511

*2002 estimated based on 3,222 incidents reported through Q3.
Note: Data from http://www.certs.org/stats/.

Command, the Federal Aviation Administration, the House of Representatives, and NATO (Fox News, 2002b).

It is not comforting knowing that the White House had been compromised rather easily, especially when a security expert was quoted as saying that "such attacks have been common for years." Equally unsettling is that most private businesses probably do a better job of protecting their sites and vital data than do many government entities. "Most private business sites probably have Web servers equipped with firewalls that are capable of halting these types of attacks. Government sites, however, are usually more vulnerable because their staffs are paid less and don't have access to the latest software and hardware defenses" (Weiss, 2001).

The Carnegie Mellon CERT Coordination Center has been tracking "vulnerabilities reported" since 1995, and these have shown exponential growth (Carnegie Mellon CERT Coordination Center, 2002). See Table 2 and Figure 3.

The attacks may not have resulted in any direct or lasting damage, but they are unsettling. One expert maintains, "The [Internet router] attack itself was invisible and without effect on Internet users. The attack on the DNS system did not noticeably degrade Internet performance." At the same time, he says that "as nations and critical infrastructure become more dependent on computer networks for their operation, new vulnerabilities are created—a massive electronic Achilles' heel" (Lewis, 2002).

If the White House can't keep hackers from shutting down their servers and the Internet is vulnerable to its critical routers being shut down, can't cyberterrorists mount that same attack on other locations in our vital information technology infrastructure? Could they cause interruptions in air traffic control, banking, or stock trading systems to the point that people will experience widespread dread, fear, or panic? If our military can't protect vital data from hackers, aren't they vulnerable to an attack by cyberterrorists that could cripple their ability to defend the country? How can that happen?

HOW CYBERTERRORISM OCCURS: CONVERGENCE OF THE PHYSICAL AND VIRTUAL WORLDS

Cyberspace is a world of combinations of ones and zeros that can represent the physical world, but are not the physical object(s). This virtual world has no appreciable mass.

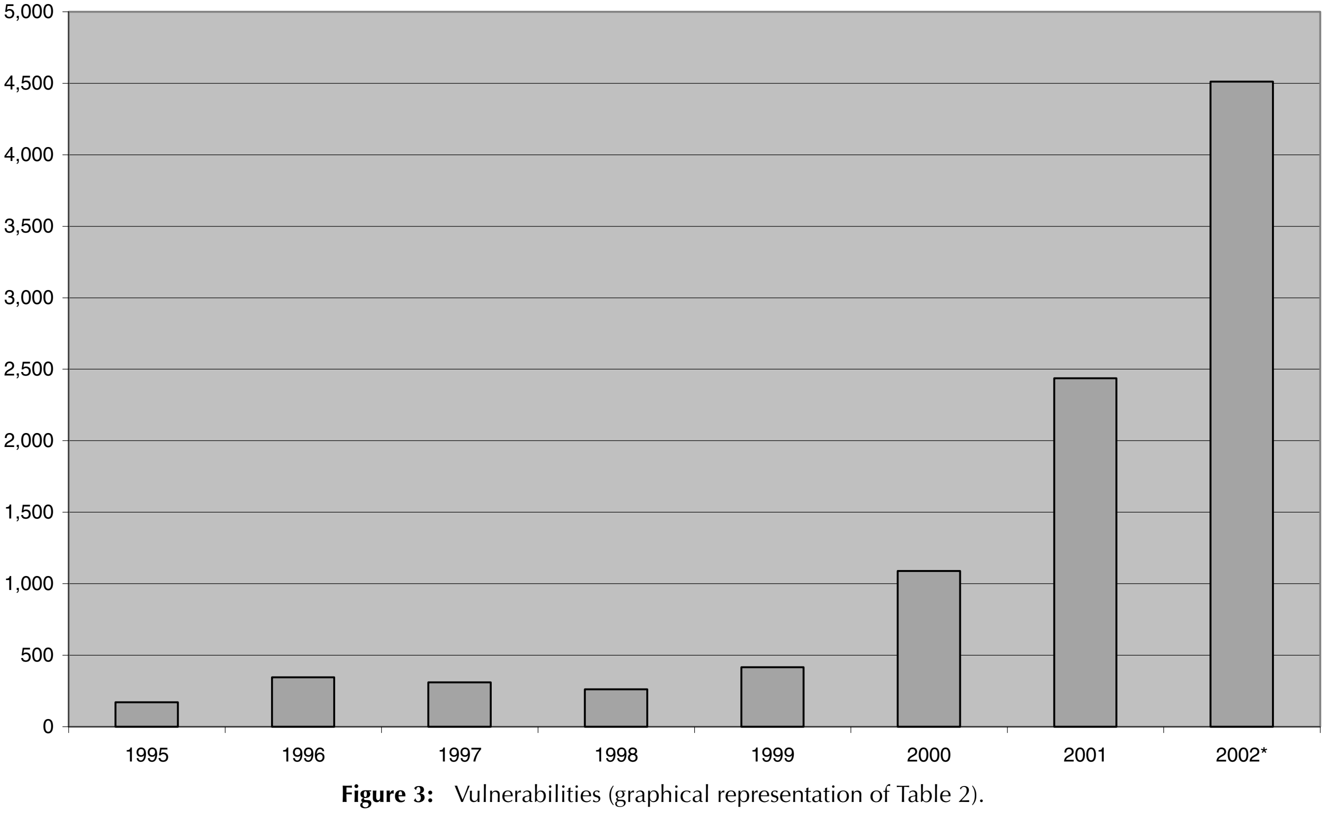


Figure 3: Vulnerabilities (graphical representation of Table 2).

It marginally exists in time and space. With the attachment of stimuli and sensory interface devices, the flow of ones and zeros can control physical objects that exist in time and space. Collin defines this interface as the convergence of the physical and virtual worlds. The physical world is matter and energy. The virtual world is symbolic, binary, and a representation of information, presumably from the physical world. "It is now the intersection, the convergence, of these two worlds that forms the vehicle of cyber-terrorism, the new weapon that we face" (Collin, 1996).

Any electronic signal can be attacked if its rule-based behavior of ones and zeros can be determined and accessed. Most of the previously cited misuse and offensive use acts use the Internet as a means to access target device points (some do not, e.g., the bombs transmitting signals leading to simultaneous detonation) through standardized Internet rules (Internet protocols, or IP, and transmission control protocols, TCP) that enable worldwide communication with devices. Each device has a unique IP address. Once terrorists have access to the IP address and the rules controlling the device itself, they have the potential to control it. Often, remote administration of the physical device is used, and control is through IP or other discoverable interfaces. These are vulnerable points.

Collin describes activities in the physical world that intersect with the cyber-based systems through these vulnerable points. His examples include food and pharmaceutical processing plants, electric and natural gas utilities, train crossings and traffic control systems, next generation air traffic control systems, virtually all modern military equipment, and military, public safety, and civilian communications. Using readily available digital tools and freely available information from Internet Web sites, the cyberterrorist exploits this "point of convergence" to achieve one or more of three goals: destruction, alteration, or acquisition and retransmission.

Denning poses several problems for the cyberterrorist, but ends up concluding that the threat is real:

> Although cyber terrorism is certainly a real possibility, for a terrorist, digital attacks have several drawbacks. Systems are complex, so controlling an attack and achieving a desired level of damage may be harder than using physical weapons. Unless people are killed or badly injured, there is also less drama and emotional appeal. . . . [Nevertheless,] some of these [computer hackers] are aligning themselves with terrorists like bin Laden. While the vast majority of hackers may be disinclined towards violence, it would only take a few to turn cyber terrorism into reality. Further, the next generation of terrorists will grow up in a digital world, with even more powerful and easy-to-use hacking tools. (Denning, 2001)

Collin might be an alarmist, but even with our best efforts at preventing cyberterrorism, some of his doomsday acts seem possible or likely.

The Tools of Cyberterrorists

Cyberterrorists are a special breed of hacker, and they use similar tools and techniques. Cyberterrorists need hardware, sophisticated software, and detailed knowledge of the technology, the targets, and the procedures used to maintain and control operations at the target sites. Much as a safecracker goes about business with torches and explosives, the cyberterrorist uses specialized cybertools to gain access and penetrate the barriers protecting targets. Hackers traditionally work alone, but that is changing. Today, hackers and cyberterrorists gain additional power by working together.

Knowledge

The cyberterrorist identifies targets and understands their intimate details. In cyberspace, it may not be necessary to know the physical location of the device, itself, but only the IP or other electronic address "where" it resides in cyberspace. Details include the physical structures in which the ones and zeros reside—usually a computer or system of computers with its associated network—and the processes used to maintain and control its action(s).

The technology to operate, control, and maintain the physical device normally has protective "firewall" barriers and is intentionally constructed to be difficult to discover and circumvent. In tracking "attack sophistication vs. intruder technical knowledge," Lipson points out that in 1980, attack sophistication was low, but it required high intruder knowledge. By 2000, attack sophistication had become high, but the required intruder knowledge was low, and there was "widespread availability of exploit tools" (Lipson, 2002).

In discussing "the democratization of hacking," John Christensen says, "The tools of mayhem are readily available. There are about 30,000 hacker-oriented sites on the Internet, bringing hacking—and terrorism—within the reach of even the technically challenged." He quotes a manager of technical security at a company that does business with the Pentagon: "It's about bragging rights for individuals and people with weird agendas . . . . You no longer have to have knowledge, you just have to have the time" (Christensen, 1999).

Software and Hardware—Cyberspace Resources

The Internet is a free-for-all. For a few thousand dollars at most, almost anyone can serve up images, information, and other content on the Web. There are no effective committees with official content authority or oversight. Government or other security agencies control information only minimally. Servers can reside in any country, sometimes in difficult to discover physical locations, and they can be moved easily.

The cost of computer hardware suitable for interacting in cyberspace is almost negligible. Better hardware enables hackers to carry out the millions of operations that may be required to discover access codes faster than they otherwise could, and may make following their trail more difficult, but almost any hardware will do.

Software is the set of instructions to control hardware. Hackers know where to find software and documentation posted on the Internet. If a hacker obtains a new tool, immediate action is needed to stop its dissemination and use. Meanwhile, new tools evolve. Such "cat and mouse" games have existed since police were organized.

Access

Operators and trustees of the target devices use increasingly sophisticated defenses, while hackers develop and share new tools and techniques to probe and defeat them. Hacking into secure sites may simply mean millions of iterations of new trial-and-error algorithms that ultimately discover the target device's secret passwords or defeat its other protective barriers. Even with low probability, the documented cases of unauthorized entry are common. An old saying is that "even a blind pig finds an acorn from time to time." In cyberterrorism, the odds of access are greater. A significant cyberterrorism incident can occur either by skill or through blind luck.

Power in Numbers

Obtaining the knowledge, hardware, software, and access to targets in cyberspace can be complex. Specialists can focus on a tiny problem that unlocks success in the broader venture. Acting together, even informally, hackers and cyberterrorists can defeat barriers that would be impossible for any single individual.

The Chinese, Pakistani, and Korean hackers were loosely associated individuals with common beliefs and an access to digital tools. Acting alone, they may have been limited to a lucky one or two successful acts. By communicating with each other on the Internet, they could exchange information about targets, hardware, software, access, timing, and target locations. They improved their probability of success by sharing. Networks of cyberterrorists working together, perhaps in widely separated parts of the world, may one day unlock the keys to making one or more of Collin's doomsday scenarios reality.

The Support Role of Cyberspace in Cyberterror

Just as businesses, governments, military forces, and other individuals use the Internet to communicate, so do cyberterrorists. It's fast, easy, cheap, and powerful. Just as the rest of us have gained power through file sharing, cyberterrorists transferred computer data to associates anywhere in the world. This is legitimate use.

They also use it for background work. Almost any computer can be accessed if left unprotected, and techniques to obtain passwords through algorithms and multiple iterations are available free on some 30,000 hacker sites (Christensen, 1999). Hackers and cyberterrorists alike use these resources to access homes, businesses, or institutions where they can plant spy devices and read the contents of computer hard drives and other data storage. The medium for cyberterror, itself, becomes the conduit for intelligence gathering. This can be use, misuse, or offensive use.

Attacks Occur through Different Countries and Different Systems

Hackers and cyberterrorists can operate out of almost any physical location on earth. Many nations have daunting problems such as hunger, disease, political unrest, or traditional war, often to the extent that they are unable to address security issues or exert control over cyberspace. This is not limited to the least developed countries. Many emerging nations have these problems and offer ample electronic infrastructure to facilitate cyberterrorism. In some cases, their political leadership is sympathetic to terrorists. In Richard Clarke's testimony before a Senate Judiciary subcommittee on cyberterrorism as cited earlier, he included Iran, Iraq, North Korea, China, Russia, and other countries as "already having people trained in Internet warfare" (Holland, 2002).

Even developed nations are vulnerable to cyberterrorism. A 1999 survey by the Victoria Police (Australia) found a third of the top 100 Australian companies had suffered from computer crime, and only 35% had a staff security awareness program. The head of the investigating squad remarked, "If a lot of [cyberterrorist] activity were to take place, Australia would be fertile ground." Some of these businesses control vital physical infrastructure or devices vulnerable to terrorism (Thom, 1999).

On October 25, 2002, the Associated Press reported that a special task force chaired by former Senators Gary Hart and Warren Rudman released a report saying that the United States remains "dangerously unprepared" to deal with another major terrorist attack, stating, "In all likelihood, the next attack will result in even greater casualties and widespread disruption to American lives and the economy" than the 9/11 attacks (AP 2002a). At about the same time, other headlines spoke of targeting railroads and voting machines (CNN, 2002b).

Vulnerabilities exist in network route traffic and in methods to keep track of users. Joel Snyder, a Canadian author, points out that the entire Internet depends on huge border gateway protocol (BGP) tables that detail the more than 100,000 routes information can travel. BGP4 is the protocol used to exchange routing information between providers and to propagate external routing information through autonomous networks. BGP-speaking routers "peer" with each other to exchange information about routes and improve data flow efficiency:

> In the early days, these tables were validated against routing registries that ensured bogus information could not be injected into the tables. Nowadays, that doesn't happen. Keeping those routing registries updated and synchronized is just too expensive and inconvenient. The lack of a global routing registry means that it's fairly easy to create routes to nowhere. . . . If a determined attacker were to start injecting routes into the BGP tables, the ripple effects could be enormous. . . . [It] could cripple routers around the world. (Snyder, 2002)

If the cyberterrorist could use BGP tables to redirect messages to random or unknown locations, it also might be possible to redirect them to specific destinations. This would be an example of acquisition and retransmission. If cyberterrorists could substitute the original path with routes to its own control station, alter the message contents, and redirect them, they may be able to take control of critical physical devices without discovering the secret passwords and other barriers to these devices! Some of Collin's dreadful cyberterror scenarios may come about through these means. While this probability is very low, vigilance is ongoing.

Groups such as the University of Oregon Route Views Project maintain servers that have regular BGP sessions with many routers spread throughout the world to study variations in core route table sizes and growth rates. Their findings are that the number of routes in the Internet increased from approximately 15,000 in 1994 to approximately 115,000 by mid-2002, magnifying their vulnerability (Meyer, 2002).

Internet domain name servers (DNS) track users, referencing millions of user names. It is common to "ping" these and other address nodes with a short query message that maintains communication with them and detects abnormalities. Normally, this traffic is minimal. The October 21, 2002, attack on 13 DNS nodes in Japan, Europe, and the United States consisted of a massive "ping flooding" denial of service attack (Naraine, 2002; Vickers, 2002). A creative terrorist might combine such an attack on the DNS servers with some form of extracting and redirecting messages through BGP routing intervention. Again, this is not likely, but it is among the many scenarios leading to potentially destructive effects.

Direct and Indirect Effects of Cyberterrorism on Vulnerable Parties

Some cyberterrorism effects—e.g., physical destruction—are obvious. Secondary results might include indirect damages to individuals, businesses, institutions, and government segments, including individual and mass dysfunction, job displacement, and other economic costs. Often, damage in one segment affects others.

Effects on Individuals

Terrifying scenes in action films, video games, novels, television shows, and other media are entertainment. Normally, they do not cause fear or dread for willing participants. H.G. Wells' famous Halloween Eve 1938 "War of the Worlds" radio broadcast created wide-spread panic, but that was not its intention (it was mistakenly believed to be real). Few people enjoy actual terror where they do not or cannot control the event.

People have experienced many instances of terror such as war; revolution; riots; bombings; the September 11, 2001, events (9/11); and the Bali bombings. Few or none have been created by cyberterrorism—yet. If disruptions of electric power, transportation, water control, or other public utilities occur through cyberterrorism, individuals are likely to react very much as they do in other terror situations, with varying degrees of fear, dysfunction, and wide-spread panic that would impact businesses, government, and society as a whole.

In the days and months following the World Trade Center attack, most individuals recovered their wits and returned to normal consumption patterns, albeit with additional caution. Consumer spending returned to near normal levels in the United States and continued to drive a recovering economy. Business activity, in contrast, was impacted, even though derived demand remained high.

Effects on Companies

Companies like stability and predictability. Spending in business markets slowed following the 9/11 attacks. Businesses were reluctant to risk economic resources in an uncertain environment. The same outcome would occur with a cyberterrorist attack. Cyberterrorism that directly disrupts plants, factories, distribution networks, or critical infrastructure lessens stability. It can drive companies out of business, with resulting economic dislocations and jobs losses.

Many businesses are aware of the potential for cyberterrorism and have taken steps to counteract it. However, an alarming number of companies are vulnerable. At a conference early in 2002, a cyberlaw expert stated, "People now have a real-world example of the issues that we discuss here. Sept. 11 has made people aware not just of the impact of loss of life, but also the impact of losing business infrastructure." Howard Schmidt, vice chairman of the Bush administration's new Critical Infrastructure Protection Board (CIP), added, "I don't think I've found anything anywhere that's free of vulnerabilities" (Moran, 2002).

Yet, many companies are not taking effective steps to protect themselves from cyberterrorists, and many are not sharing what they learn with other companies or the authorities charged with helping them. An April 2002 CNN.com article about the seventh annual Computer Crime and Security Survey, reported by the FBI and the Computer Security Institute (CSI), states, "About 90 percent of the respondents detected a security breach within the last 12 months. However, only 34 percent reported the intrusions to law enforcement officials." The CSI director added, "There is much more illegal and unauthorized activity going on in cyberspace than corporations admit to their clients, stockholders and business partners or report to law enforcement." Bruce Gebhardt, executive assistant director of the FBI said, "Now, more than ever, the government and private sector need to work together to share information and be more cognitive of information security so that our nation's critical infrastructures are protected from cyber-terrorists" (Sieberg, 2002). Why are companies reluctant to cooperate?

Companies want to avoid the bad publicity associated with having had their systems and facilities compromised by a cyberterrorist, and they want to protect their proprietary information from scrutiny by agencies that would turn such information into public record. In many cases, from a purely business point of view, it makes more sense to deal with the issue internally and not report it to police or other investigative agencies that might leak or subject it to the Freedom of Information Act. "Corporate leaders, in many instances, simply never tell the outside world they've been victimized, to avoid spooking investors or customers" (O'Connor, 2000).

Effects on Society

The general public has a limited understanding of cyberspace, the extent to which our infrastructure is interconnected by cyberspace, or the potential devastation that could occur by controlling air traffic, railroad switches, dam spillways, nuclear power, and other powerful devices. Although there is almost no research about cyberterrorism's effects on society, it is reasonable to assume that people will behave similarly to the way they react to conventional terrorism.

Society is composed of individuals and institutions. Many individuals become dysfunctional in the face of terror. They adopt negative attitudes and experience loss of morale. If society at a macro level becomes demoralized, traumatized, panicked, or filled with dread, as a successful cyberterrorist attack might accomplish, the collective synergy would have the potential to bring about a difficult-to-control negative spiral affecting financial institutions, securities markets, and individuals' financial conditions. Many individuals never fully recover from the psychological or financial effects of a terror attack.

However, if people perceive a terror source to be external and hostile to their way of life, it can bring society together with a common purpose. This occurred following the 9/11 attacks in the United States and the Bali attacks in Australia. The causes of traditional terrorist attacks are easily seen through on-the-spot media reporting, and many individuals throughout the world came together to condemn the terrorists. A cyberterror act might be more difficult to understand, and thus, the results are uncertain.

Effects on Government

Most governments ultimately govern through the support of their interrelated individuals, businesses, and societal entities. Successful cyberterrorism would weaken each supporting element, but the common purpose scenario might benefit a government and make it stronger, which occurred immediately following the 9/11 attacks. Much depends on how the government handles the crisis.

Effective government services include infrastructure development and maintenance, communication, defense, and other functions that could be directly affected by a successful cyberterrorist attack. If not brought under control, they could lead to a loss of confidence, ability to maintain order, and, potentially, anarchy and chaos.

Governments around the world have leveraged world terrorism and galvanized concern into support for strong antiterrorist measures. In the United States this has included aggressive new laws allowing an expansion of wiretapping and other surveillance activities. The USA Patriot Act of 2001 broadly expands law enforcement's surveillance and investigative powers. In December 2002, FBI Director Robert Mueller reported that since the World Trade Center attack, "tens of attacks, probably close to a hundred around the world" have been detected and stopped (AP, 2002c).

WHAT TO DO TO ELIMINATE OR MINIMIZE CYBERTERRORISM?

A long struggle with cyberterrorism may be just beginning. It could be a battle of minimization, not elimination. In the days following the 9/11 attacks, Secretary of Defense Donald Rumsfeld said,

> The cold war, it took 50 years, plus or minus. It did not involve major battles. It involved continuous pressure. It involved cooperation by a host of nations. And when it ended, it ended not with a bang, but through internal collapse. It strikes me that might be a more appropriate way to think about what we are up against here. (World Almanac, 2001)

In his appearance before the Senate judiciary committee previously mentioned, Richard Clarke said, "If I was a betting man, I'd bet that many of our key infrastructure systems already have been penetrated.... We reserve the right to respond in any way appropriate: through covert action, through military action, any one of the tools available to the president" (Holland, 2002). Most individuals have resigned themselves to some loss of freedoms to counteract the threats.

Individuals, businesses, institutions, and society are working to eliminate or minimize the effects of cyberterrorism. Individuals are turning off their computer at night and isolating their computers from networks, either through software or by physically disconnecting them. Each such defensive action carries its own trade-offs of individual freedom. Private organizations, higher education institutions, and conferences are contributing, often before the government is able to respond. Collectively, this helps mitigate the threat of cyberterrorism.

Addressing the Threat at Every Level

Government Branches and Entities

At Stanford University, on September 18, 2002, the United States released a draft of the National Strategy to Secure Cyberspace. It included 60 recommendations for government, companies, institutions, and individuals, including those in other nations, to help promote cybersecurity and prevent cyberterrorism. Six categories of defense included education, training, company procedures, new security technologies, federal modeling, and early warning and crisis management (Porteus, 2002b).

Critics on one side believe that the recommendations are not forceful enough and argue for more laws and regulations. Part of the strategy addresses "administration concerns that efforts to secure cyberspace are hampered by the lack of a single data-collection point to detect cybersecurity incidents and issue warnings" (Carlson & Fisher, 2002). Mark Rosh, formerly the Justice Department's top computer crimes prosecutor, said, "All of these are good recommendations, but none have the force of law. There is no carrot and there is no stick. You need to put some teeth into some of the proposals" (CNN.com, 2002a). Others believe that they are too invasive and infringe on civil liberties and privacy.

The federal government has a delicate balancing act in attacking terrorism and cyberterrorism, while maintaining personal liberties, freedoms, and business creativity and flexibility. It is probably impossible to satisfy both sides. Other nations have similar issues.

The U.S. government has established the Department of Homeland Security, centralizing many government agencies and branches into a more unified voice for establishing effective policies to help protect citizens and the government infrastructure against cyberterrorism. There have been ongoing disagreements about the specific implementation of this merged agency, but not the broad concept or belief that such an agency can help combat the threat.

Inside or outside the Department of Homeland Security, the federal government operates numerous intelligence gathering agencies, including the FBI, CIA, and supersecret units. For several years, they were constrained in what they could do with respect to human intelligence, actual agents on the street. In response, they implemented surveillance technologies that can monitor worldwide telephone conversations, e-mail, and other confidential correspondence. Such technologies can be very useful in intercepting and monitoring cyberterrorist threats, but they raise substantial privacy concerns. With the 9/11 attacks and increased terrorist activity worldwide, public opinion, balancing self-preservation with other needs, has swung to cautious support.

The National Strategy to Secure Cyberspace is intended to serve as a framework for federal action and as recommendations for companies, groups, and individuals. The strategy "strives to ensure that any interruptions will be infrequent, brief, manageable, geographically isolated, and minimally detrimental to the welfare of the United States" (Porteus, 2002a). Specific recommendations for government include improving federal cybersecurity so it can be a model for accountability, developing its own early warning and crisis management plans, and helping to promote science and industry in developing new security technologies.

Companies

The National Strategy to Secure Cyberspace business sector strategy calls for a heightened awareness and responsibility within companies. It promotes company-wide corporate security councils to integrate all aspects of security, including cybersecurity. Companies have much to lose through cyberattacks, but there is wide variation in how individual companies have taken action to protect themselves from various cyberspace attacks.

Some companies have been lax in applying even modest protective measures. In October 2002, the Associated Press reported on the W.32.Bugbear virus, calling it "the worst computer security outbreak in the world" (AP, 2002b). Once a computer is infected, the hacker can steal and delete information. The worm is expected to last well into 2003 because many users will not realize that their computer is infected. Incidents such as these, while destructive in the short term, will have a beneficial effect in prodding laggard businesses.

There is wide agreement that companies should do a better job of reporting cases of unauthorized access, cybercrimes, and potential cyberterrorism. The strategy is an attempt to raise awareness and spur them to action in both reporting and fixing security breeches. "The expectation is that shareholders will eventually hold boards accountable for security breaches, and, in turn, boards will hold security officials responsible" (Carlson, 2002).

Private "Pinkertons of Cyberspace" Firms

In the past several years, an entire service industry has grown up around the idea of helping companies with their cybersecurity needs. From a business point of view, creating an internal unit for this purpose may be costly, require the recruitment of specialized personnel, entail substantial start-up time, and risk becoming isolated from the latest developments. For many companies, it makes more sense to outsource their cybersecurity needs. Many service providers have highly trained people who are able to leverage their information and knowledge over several businesses.

Companies are investing in new security measures through software, private networks (only minimally connected to the Internet), and services. "Sales for firewalls and virtual private networks (VPNs) should hit $7.5 billion in 2005 and e-security services are expected to increase about 24 percent a year between 2001 and 2005. . . . Global investment in e-security services should total $14.5 billion by 2005. . . . North America will account for 58 percent of the global security product market this year" (McMahon, 2002). The world will follow.

Institutions

Nonprofits and private and quasi-private institutions typically do not have the same security needs as private companies. In many cases, they hold considerable financial and information resources that need protection, but it would be rare for them to have control over nuclear power plants, electrical grids, dams, or other vital infrastructure. Many of the techniques companies are using apply equally to institutions. However, institutions seldom have enough in-house expertise to combat potential cyberterrorism. They are good candidates for private security firms, but resource constraints make it more difficult to afford the needed services. This is an area that will benefit from closer attention.

Individuals and Society

In the past several years, the number of individuals engaged in cyberspace has skyrocketed in the developed world. E-mail is the most common application, but many users have become intensive participants in chat rooms, instant messaging, music and video swapping, and finding and contributing other text, graphic, and multimedia Web content. These societies have migrated to the information age. Many individuals and organized groups have become active in supporting new antispam laws, and many have taken measures to protect their computers from unauthorized access.

Most individuals become exposed to cybercriminals—not cyberterrorists—through credit card and identity theft or access confidential address books and other computer files involving property. Unscrupulous businesses send incessant e-mail (spam) and refuse to abide by good conduct norms. The "I Love You" virus caused an estimated $4 billion in damages to individuals and international systems (IWDB, 2000). Occasionally, individuals are

Table 3 Number of Security Notes

Year	1998	1999	2000	2001	2002*
Incident Notes	7	8	10	15	7
Vulnerability Notes	8	3	47	326	445

*2002 estimated based on 5 and 318 Notes reported through Q3.
Note: Data from http://www.cert.org/stats/.

exposed to bodily harm, as in the Florida cybersex case, but these are cyberterrorism only in the loosest sense. Disabling private computers likely would be a minor part of a cyberterrorist's plan.

The National Strategy to Secure Cyberspace makes recommendations to private users to regularly update antivirus systems, turn off computers at night, and install firewall software. The need for these measures has been a significant deterrent to expansion of broadband 24/7 connections. In addition, the strategy recommends that users apply caution in opening e-mail attachments that may contain viruses, worms, or other invasive components.

Organizations and Conferences

Many private and semiprivate groups combat cybercrime and cyberterrorism. For example, the Carnegie Mellon Software Engineering Institute's CERT Coordination Center, which tracks cyberspace "incidents" and "vulnerabilities," publishes a stream of security notes designed to help businesses, organizations, educational institutions, government, and technical individuals deal with incidents and vulnerabilities. Vulnerability notes rose dramatically beginning in 2001 (Carnegie Mellon CERT Coordination Center, 2002). See Table 3 and Figure 4.

The SANS Institute (SysAdmin, Audit, Network, Security) in conjunction with the FBI publishes a Top 20 list of security threats segmented by the two large and vulnerable operating systems used in corporate networks—Windows and UNIX. The list is "especially intended for those organizations that lack the resources to train, or those without technically advanced security administrators" (Wagner, 2002).

The National Intelligence Council publishes papers on global trends under the aegis of the National Foreign Intelligence Board and the Director of Central Intelligence, sometimes together with other government or private centers, and sponsors conferences such as the Future Threat Technologies Symposium and The Global Course of the Information Revolution: Technological Trends (NIC, 2000).

The National Infrastructure Protection Center (NIPC) and the National Association of State Chief Information Officers (NASCIO) have partnered to form an Interstate Information Sharing and Analysis Center (ISAC) "to disseminate intelligence quickly and prevent unauthorized and destructive infiltrations" (Greenspan, 2002b).

Some of these organizations are also in business providing services to commercial, nonprofit, government, and other entities. Many participate in, sponsor, or promote conferences calling attention to cyberterrorism issues and cybercrimes, and are working actively with the Bush administration on implementing the National Strategy to Secure Cyberspace. Many groups are working with the Department of Homeland Security.

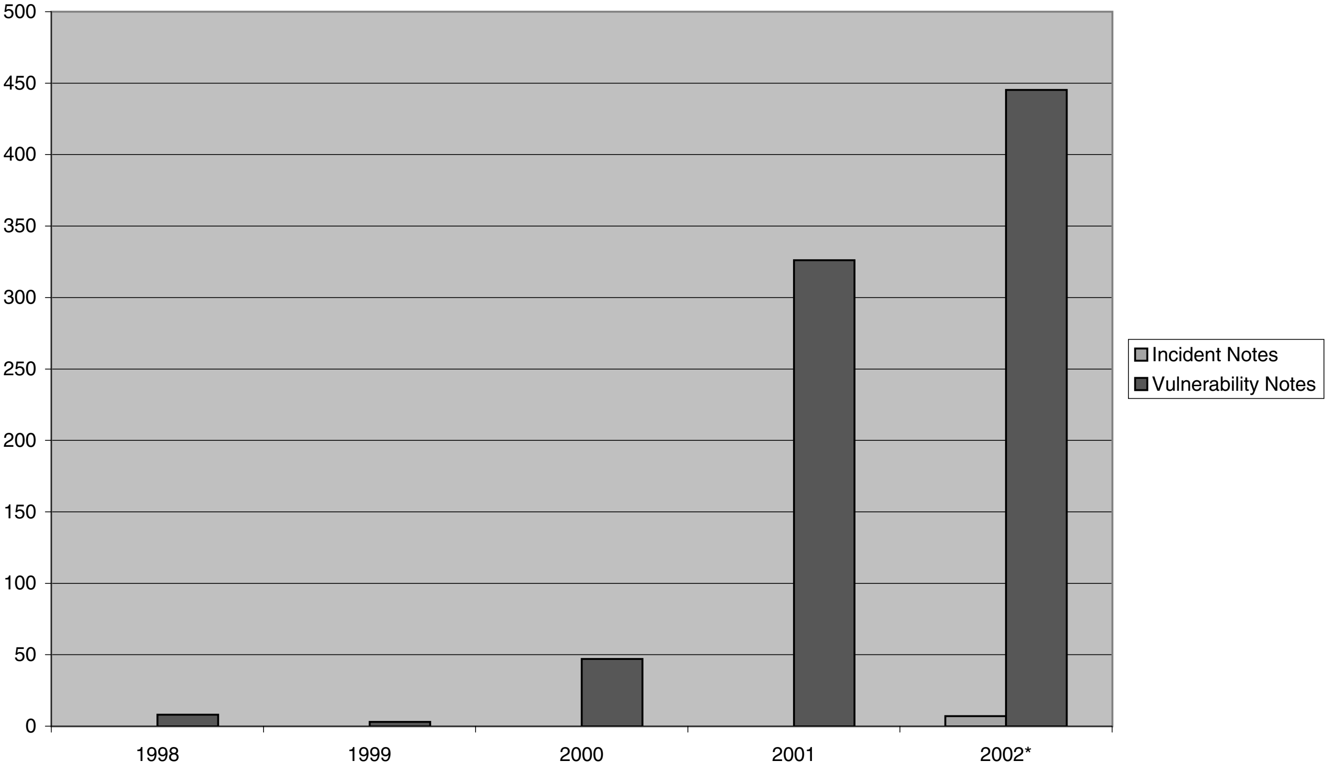


Figure 4: Security notes (graphical representation of Table 3).

Higher Education Institutions

In September 2002, it was reported that Congress approved National Science Foundation (NSF) grants for seven large projects and 240 smaller projects that will disburse between $500,000 and $13.5 million to each recipient organization, many of which are higher education institutions (Legon, 2002b).

Colleges, universities, and other higher education institutions were pioneers in using and promoting the Internet. Long before the graphical World Wide Web existed and before businesses were allowed to use the Internet for commercial purposes, the academic community was active in file sharing and facilitating other group research through the Internet. Many higher education institutions became leaders in developing security skills and experience with the UNIX operating system and its associated Internet protocols, TCP/IP. Vigilant individuals in higher education have a tradition of working long and hard—often with little or no extra compensation—in making UNIX and the Internet as robust as it is today.

Today, in community colleges, public and private universities, and specialized institutions such as the American Military University, innovative technologies are being developed and new courses are being created and taught about how to counteract terrorism in all its forms, including cyberterrorism. Many of these are overflowing. In a letter to alumni, John Hennessy, President of Stanford University, said, "Professor William Perry, former Secretary of Defense, was one of the many faculty members who made extraordinary efforts to accommodate students in oversubscribed classes. Bill's class on 'Technology in National Security' swelled from an enrollment of 145 in the fall of 2000 to 329 students this past year!" (Hennessy, 2002).

Legal and Privacy Concerns

With the rush to fight back against cyberterrorism and other cyberspace crimes and vandalism, a small but vocal group in the United States continues to maintain that these ends do not justify their means. They argue that the new laws giving law enforcement greater powers are acts of desperation and violate fundamental rights guaranteed by the Constitution and the Bill of Rights. In their view, the resulting loss of freedom is not worth the security benefits.

In January 1999, Ed Gillespie, Executive Director of Americans for Computer Privacy, attacked President Clinton's policies restricting the dissemination of encryption software. "This [protection of IT infrastructure] job can best be done by sophisticated encryption technology produced by U.S. high-tech companies. [It is] essential to fighting cyber-terrorism" (ACP Press Release, 1999).

Others argue that terrorists are helped by encryption products if the government is subsequently unable to monitor their communications. "Terrorists may be tech-savvy as well. They often use encryption programs when communicating over the Internet, scrambling the text of messages as they travel over the Web." The government fights terrorism with programs such as Carnivore and Magic Lantern, which "allow investigators to monitor personal e-mails and electronic communications." Other government agencies and private companies are making software to monitor file-trading activity, downloads, and message board notes with suspicious content. Strong encryption would hamper many of these efforts (Porteus, 2002a).

The new strategy, new laws, and new technology enable government agencies to seek out and identify terrorists. They also increase government and companies' ability (and perhaps willingness) to collect personal data. Fred Cohen, the keynote speaker at a three-day cyberterrorism conference in Connecticut, urged attendees to avoid many of the proposed security measures. He cited a potential national database on citizens and increased government surveillance: "We're doing the wrong stuff. It's not effective, and it's oppressive" (Moran, 2002).

Former Rep. Bob Barr, R-Ga., working with the American Civil Liberties Union, agrees:

> The only way that you can make this sort of system work, that they are talking about, is if you have access to all, virtually all the e-mail traffic, all credit card transactions, all medical records, all gun purchases. Otherwise the sort of system they are trying to develop here, where they can do cross-referencing and develop profiles, won't have any meaning. (Fox News, 2002a)

Cohen and Barr's concerns may be moot. In an interview with the *Boston Globe*, Richard Hunter discussed companies' data-collection policies and excerpts from his book, *World Without Secrets*: "Information is constantly recorded and made available to almost anyone who wants it, regardless of intent.... The amount of information that's out there, and readable, is already huge" (Denison, 2002). Companies already have most of the data, the government can get it, and in the wake of the 9/11 attacks and continuing terrorist threats, public opinion has become less sensitive to private data collection.

International Law Enforcement

What the United States and other developed nations are doing to protect their citizens can be applied on a global scale. However, the nations of the world are struggling to cooperate on international justice issues. So far at least, beyond prosecuting international war criminals charged with physical atrocities, such efforts have made only minimal progress.

International law and bilateral treaties rarely adjudicate actions directly concerning individuals. Entities such as the World Trade Organization set rules for engaging in trade and commerce and have a judicial branch to settle differences between nations in trade matters. They do not deal with conflicts between companies, individual entities, or public–private disputes. International law and treaties usually deal only with physical presence and events, but activities in cyberspace do not directly trespass.

Nations are reluctant to give up their sovereignty in prosecuting crimes or acts of terrorism to an international body. Many developed nations have set up their own international security agencies, many of which operate anywhere in the world—and often in the virtual world. Nations with such agencies generally work with each other through informal agreements. Since the mid-1990s, cooperation has been on the upswing, and the 9/11 attacks and other international terrorism incidents have added urgency. Although it is rarely discussed publicly, this is a very active area of international relations.

CONCLUSION

Terrorism, which has been a fact throughout human history in one form or another, has found a powerful new weapon. Operating in cyberspace and enabled with today's powerful technology, the cyberterrorist may strike from anywhere, anytime, and be very difficult to trace and bring to justice.

Cyberterrorism is in its early stages. Some experts argue that terrorists are just now beginning to learn the skills to turn cyberspace into perhaps the world's most effective weapon for asymmetric responses—and terror! Others say the threat is overblown, and governments, business, organizations, and other stakeholders will take effective measures to prevent its occurrence.

Very likely, the truth is somewhere in between. The battle with cyberterrorism is expected to be long, but it will not lead to doomsday scenarios. It is unlikely that the civilized world will be able to eliminate it, but it may be controlled. People will need to learn to live with cyberterrorism and minimize its impact.

This chapter overlaps with many others in this volume, including chapters about cyberlaw, legal issues, cybercrime, cyberstalking, privacy issues, and the organizational impact of the Internet and e-commerce. The reader should consult these other chapters for a fuller understanding of cyberterrorism.

GLOSSARY

9/11 The World Trade Center and Pentagon attacks, September 11, 2001.

24/7 An activity that can occur 24 hours per day 7 days per week, as do many broadband Internet connections.

Application software Software that allows a user to apply a hardware device to a task or series of tasks, such as word processing or Web browsing.

Asymmetric response A one-sided response and a generic explanation why cyberterrorists behave as they do. Unable to achieve their objectives through conventional means, an extremist may strike against the "enemy" with a one-sided act of terror.

Border gateway protocol (BGP) Tables that detail the more than 100,000 routes information can travel over the Internet. BGP routers exchange information about routes and improve data flow efficiency.

Cyber- A prefix for topics related to computers and/or networks, including acts of terror using computers and/or networks.

Cybercrime Activity in cyberspace usually associated with fraud, commercial espionage, or theft of intellectual property, identity, or private data, and can result in loss of data or corruption of computer files.

Cyberprotests Short periods of intense political activity in furtherance of political or social objectives.

Cyberspace, digital world, virtual world The world of discrete mathematical values, usually ones and zeros at its lowest level, that powers computers and networks and can be used to represent the physical world. It is generally outside time and space, unless it is applied to a physical object through an interface to the analog world.

Cyberterrorism An evolving body of activity that includes both acts *in* cyberspace and *using* cyberspace tools to create fear or panic, often done by subnational groups or clandestine agents and directed toward intimidating or coercing a government or some segment of the noncombatant civilian population, usually in furtherance of political or social objectives; differs from hacking and cybervandalism in its intent and the degree of severity of the attacks.

Denial of service A condition under which a Web site or other Internet resource is disabled by an attack from an overwhelming number of inbound messages.

Dysfunctional A state in which a person, object, or device has some degree of functional impairment, and can be caused by cyberterrorism.

Extremist or fanatic terrorist A terrorist that attracts a small following. Often, their methods are widely condemned, or at least not supported by other parties who may oppose the target of the terrorist.

Government Federal, state, local, military, police, justice, legislative, executive, and related entities.

Graphical user interface (GUI) Typically used to describe the graphical interface between a computer's display and the internal hardware workings, or the interface between a computer's display and images served up on the Web.

Hacker A person who gains unauthorized entry into computers or other digital devices through cyberspace, usually for the purpose of minor damage or annoyance, such as defacement of a Web site or observation of data.

Hacktivism Generally a long-term series of cyberattacks in furtherance of political or social objectives.

Hardware Physical devices that process mathematical ones and zeros. Hardware has an interface to the physical world, such as a keyboard or monitor display.

Internet The worldwide connection of computing nodes and its associated network that uses Internet protocols and transmission control protocols.

Information technology (IT) The general use of cyberspace for communicating and passing data.

Operating system Software that tells a hardware device how to operate, or run itself.

Physical world, analog world The world of physical objects and forces in time and space. The analog world is continuous. It can be represented in cyberspace by devices that use digital ones and zeros.

Software Programming code that, at its basic level, delivers ones and zeros that can be used to control the operation and function of hardware devices.

Transmission control protocols/Internet protocols (TCP/IP) The set of rules by which the Internet operates and communicates between devices.

Terrorist A person or group that usually believes strongly in their cause and is willing to engage in destructive acts that leave a trail of destruction, fear, dread, or panic to communicate their message.

Web browser Software that allows a computer to display graphical images that have been served up on the Web.

Web server Software that allows a computer to formulate, represent, and present graphical images on the Web such that they may be displayed by a Web browser.

World Wide Web (a.k.a. Web) The graphical overlay that resides on top of the Internet and allows graphics to be served and browsed.

CROSS REFERENCES

See *Computer Viruses and Worms; Cybercrime and Cyberfraud; Cyberlaw: The Major Areas, Development, and Provisions; Cyberstalking; Denial of Service Attacks; International Cyberlaw; Internet Censorship; Legal, Social and Ethical Issues; Privacy Law.*

REFERENCES

ABCNews.com (2001, October 1). *What kind of war? Transcript: Military analyst Tony Cordesman on America's new war*. Retrieved December 23, 2002, from http://abcnews.go.com/sections/community/DailyNews/chat_cordesman1001.html

ACP Press Release (1999, January 22). *ACP statement on administration's cyberterrorism initiatives* Washington, DC: Americans for Computer Privacy.

Annan, K. (1999). *Report of the Secretary General on the work of the Organization* (A/54/1/, para. 254). Retrieved from http://www.un.org/Docs/SG/Report99/toc.htm

Associated Press (AP) (2002a, October 25). Report: Next attack could top 9/11. Retrieved October 25, 2002, from http://www.foxnews.com/story/0,2933,66668,00.html

Associated Press (AP) (2002b, October 7). Stealthy e-mail worm bores into computers in a dozen countries. *Medford Mail Tribune*.

Associated Press (AP) (2002c, December 16). FBI Director: 100 terror attacks thwarted. Retrieved December 16, 2002, from http://www.foxnews.com/story/0,2933,73122,00.html

Carlson, C. (2002, September 9). Feds pitch cyber-fence. *eWeek, 19*(36), 18.

Carlson, C., & Fisher, D. (2002, August 26). Bush to call for fed NOC. *eWeek, 19*(34), 1.

Carnegie Mellon CERT Coordination Center (2002). *CERT/CC statistics 1988–2002*. Retrieved January 8, 2003, from http://www.cert.org/stats

Christensen, J. (1999, April 6). Bracing for guerrilla warfare in cyberspace. Retrieved July 25, 2002, from http://www.cnn.com/TECH/specials/hackers/cyberterror/

Clinton, W. (2000, December 11). A foreign policy for the global age. Address at the University of Nebraska. Retrieved August 14, 2002, from http://usembassy.state.gov/islamabad/wwwh00121101.html

CNN.com (2002a, September 18). Cybersecurity plan avoids call for new rules. Retrieved September 18, 2002, from http://www.con.com/2002/TECH/internet/09/18/cybersecurity.ap/index.html

CNN.com (2002b, October 25). FBI: Al Qaeda operatives may target U.S. railroads. Retrieved October 25, 2002, from http://www.cnn.com/2002/US/10/25/railroad.warning/index.html

CNN.com (2002c, July 5). Yale accuses Princeton of hacking. Retrieved July 25, 2002, from http://www.cnn.com

CNN.com (2002d, August 23). FBI raids firm after military hacking claim. Retrieved August 23, 2002, from http://www.cnn.com/2002/TECH/internet/08/23/computer.security.ap/index.html

Collin, B. C. (1996). The future of cyberterrorism: Where the physical and virtual worlds converge. Paper presented at 11th Annual International Symposium on Criminal Justice Issues. Retrieved July 23, 2002, from www.afgen.com/terrorism1.html

Convention on Cybercrime (2001a). 30 states sign the Convention on Cybercrime at the opening ceremony. Retrieved December 23, 2002, from http://press.coe.int/cp/2001/875a(2001).htm

Convention on Cybercrime (2001b). The Convention on Cybercrime, a unique instrument for international cooperation. Retrieved December 23, 2002, from http://press.coe.int/cp/2001/893a(2001).htm

Conway, M. (2002, November). Reality bytes: Cyberterrorism and terrorist "use" of the Internet. *First Monday, 7*(11). Retrieved December 20, 2002, from http://firstmonday.org/issues/issue7_11/conway

Cyberatlas staff (2001, December 13). Internet, computer security concerns Americans. Retrieved July 23, 2002, from http://cyberatlas.internet.com/big_picture/geographics/article/0,,5911_939161,00.html

DCI Counterterrorist Center, CIA (2002). The war on terrorism: A call to action. Retrieved August 14, 2002, from http://www.cia.gov/terrorism/ctc.html

Denison, D. C. (2002, May 19). "Smart" stats: business at the speed of a fastball. *The Boston Globe*.

Denning, D. (2001, November 1). Is cyber terror next? Retrieved December 20, 2002, from the Social Science Research Council Web site: http://www.ssrc.org/sept11/essays/denning.htm

FBI (1999). News Release, Tampa, FL 33602, November 10, 1999. Retrieved August 6, 2002, from http://tampa.fbi.gov/pressrel/1999/11_10_99.htm\

FOX News (2002a, November 27). Barr to join ACLU. Retrieved November 27, 2002, from http://www.foxnews.com/story/0,2933,71553,00.html

FOX News (2002b, December 6). Feds raid software company suspected of terror ties. Retrieved December 6, 2002, from http://www.foxnews.com/story/0,2933,72345,00.html

Greenspan, R. (2002a, September 27). Computers still insecure. Retrieved September 28, 2002, from http://cyberatlas.internet.com/big_picture/applications/article/0,1323,1301_1472111,00.html

Greenspan, R. (2002b, August 16). Cyberterrorism concerns IT pros. Retrieved September 28, 2002, from http://cyberatlas.internet.com/big_picture/geographics/article/0,,5911_1448291,00.html

Hennessy, H. L. (2002, August). Letter to alumni, Stanford University.

Holland, J. (2002, February 13). White House expert says US may retaliate with military if terrorists try cyberterrorism. Associated Press. Retrieved July 31, 2002, from LEXIS-NEXIS.

Information Warfare Database (IWDB) (2002). The Terrorism Research Center, in cooperation with Georgetown University. Retrieved October 2, 2002, from http://www.terrorism.com/

Jane's Information Group Ltd. (1999, October 21). "China-Taiwan hacker wars," *000/2565*. Retrieved August 14, 2002, from http://www.infowar.com/hacker/99/hack_10299a_j.shtml;Internet

Johnson, M., and Radcliff, D. (2001, November 9). Cybersecurity czar: Protect IT infrastructure. Retrieved July 31, 2002, from http://www.cnn.com/2001/TECH/internet/11/09/infrastructure.protection.idg/index.html?related

Legon, Jeordan (2002b, September 27). New net project aims to avoid hacking. Retrieved September 28, 2002, from http://edition.cnn.com/2002/TECH/internet/09/27/iris.internet/

Lewis, J. A. (2002, December). *Assessing the risk of cyber terrorism, cyber war and other cyber threats* Washington DC: Center for Strategic & International Studies. Retrieved from http://www.csis.org/tech/0211_lewis.pdf

Lipson, H. F. (2002, November). Tracking and tracing cyber-attacks: Technical challenges and global policy issues. Retrieved December 23, 2002, from http://www.cert.org/archive/pdf/02sr009.pdf

McMahon, T. (2002, June 13). Terrorist attacks mean big e-security spending. Retrieved December 23, 2002, from http://www.europemedia.net/shownews.asp?ArticleID=10960

Meyer, D. (2002). *University of Oregon Route Views Project*. Retrieved December 23, 2002, from the Advanced Network Technology Center Web site: http://www.antc.uoregon.edu/route-views

Meriweather, D. (1995). *Takedown*. Retrieved August 14, 2002, from http://www.takedown.com

Moran, J. M. (2002, February 5). Eye on cyberterrorism. *The Hartford Courant*.

Naraine, R. (2002, October 23). Massive DDoS attack hit DNS root servers. Retrieved October 25, 2002, from http://www.internetnews.com/ent-news/article.php/1486981

NIC (1999, October 14). Buck Rogers or rock throwers? Conference report. Retrieved August 14, 2002, from http://www.cia.gov/nic/pubs/conferenece_reports/buck_rogers.htm

NIC (2000, December). Global Trends 2015: A dialogue about the future with nongovernment experts. Retrieved August 14, 2002, from http://www.cia.gov/nic/pubs/2015_files/2015.htm

NIPC (2001a, October). CyberProtests: The threat to the U.S. information infrastructure. National Infrastructure Protection Center. Retrieved July 25, 2002, from http://www.nipc.gov/

NIPC (2001b, June 15). Cyberterrorism: An evolving concept. National Infrastructure Protection Center. Retrieved July 25, 2002, from http://www.nipc.gov/

O'Connor, R. (2000, November 6). Cracker jacked! Feds losing battle against cyberintruders. *Interactive Week, 7*(45), 14–16.

Park, K. (2001). *World almanac and book of facts*. New York: World Almanac Books.

Polit, M. M. (2002). Cyberterrorism—Fact or fancy? Retrieved August 14, 2002, from http://www.cs.georgetown.edu/~denning/infosec/pollitt.html

Porteus, L. (2002a, October 13). Feds use high-tech spyware to nab terrorists. Retrieved October 14, 2002, from http://www.foxnews.com/story/0,2933,65528,00.html

Porteus, L. (2002b). White House releases cyber-security plan. Retrieved September 19, 2002, from http://www.foxnews.com/story/0,2933,63452,00.html

Raines, P. S. (2001, April 1). Virus innoculation: ISPs should scan e-mails for viruses. Retrieved July 25, 2002, from http://www.softwaremag.com/archive/2001/apr/PRaines.html

Regan, T. (1999, July 1). When terrorists turn to the Internet. *Christian Science Monitor,* p. 17. Retrieved August 14, 2002, from http://csmweb2.emcweb.com/durable/1999/07/01/p17s1.htm

Sargent, M. (2001). Twisted list: Five most notorious hackers ever. Retrieved August 14, 2002, from http://www.techtv.com/screensavers/twistedlist/story/0,24330,3321221,00.html

Sieberg, D. (2002, April 7). Cybercrime rising, yet fewer companies reporting incidents. Retrieved July 31, 2002, from http://www.cnn.com/2002/TECH/internet/04/07/cybercrime.survey/index.html?related

Snyder, J. (2002, June 28). Could the next terror target be the Internet? *Canadian Business and Current Affairs*. Retrieved July 31, 2002, from LEXIS-NEXIS.

Sproles, J., & Byars, W. (1998). Cyber-terrorism. Retrieved August 14, 2002, from http://www-cs.etsu-tn.edu/gotterbarn/stdntppr/

Tang, R. (2001, May 1), "China–U.S. cyber war escalates." Retrieved August 14, 2002, from http://www.cnn.com/2001/WORLD/asiapcf/east/04/27/china.hackers

Thom, G. (1999, July 7). Web of fear—Cyberterror may be the price we pay for the growth of the Internet. *Herald Sun*. Retrieved August 14, 2002, from http://www.infowar.com/class_3/99/class3_080299a_j.shtml

Vatis, M. (2001). Interview with Michael Vatis, Chief of National Infrastructure Protection Center (NIPC) [Television Broadcast]. New York and Washington: Public Broadcasting Service. Retrieved July 25, 2002, from http://www.pbs.org/wgbh/pages/frontline/shows/hackers/interviews/vatis.html

Vickers, A. (2002, October 25). FBI hunt megahackers who blitzed Internet. *The Mirror*. Retrieved November 7, 2002, from LEXIS-NEXIS.

Wagner, J. (2002). SANS/FBI names top 20 network threats. Retrieved October 2, 2002, from http://www.internetnews.com dev-news/article.php/1474281

Watson, D. L. (2002). The terrorist threat confronting the United States. Retrieved August 6, 2002, from http://www.fbi.gov/congress/congress02/watson020602.htm

Webster's new world dictionary of the American language encyclopedic edition (1952). Friend, J., and Guralnik, D. (Eds.). Cleveland and New York: World Publishing Company.

Weiss, T. R. (2001). Denial-of-service warning issued by FBI. Retrieved July 31, 2002, from http://www.cnn.com/2001/TECH/internet/05/08/dos.warning.idg/index.html

FURTHER READING

Legon, J. (2002a, October 23). FBI seeks to trace massive Net attack. Retrieved October 24, 2002, from http://www.cnn.com/2002/TECH/internet/10/23/net.attack/index.html

Pastore, M. (2001, April 17). Cybercrime worries Americans, with good reason. Retrieved July 23, 2002, from http://cyberatlas.internet.com/big_picture/hardware/article/0,1323,5921_744811,00.html

Springer, D. (2002). Electronic voting worries add to booth tensions. Retrieved October 24, 2002, from http://www.foxnews.com/story/0,2933,66562,00.html

D

Databases on the Web

A. Neil Yerkey, *University at Buffalo*

INTRODUCTION

Requisite technologies have converged sufficiently to allow dynamic access to databases using Web browsers. The Web has provided a global infrastructure, a set of standards, and a presentation format, while database technology has contributed storage techniques, query languages, efficient access to large bodies of highly structured data, and mechanisms for maintaining the integrity and consistency of data (Abitebol, Buneman, & Suciu, 2000).

Prior to the Web, Internet access to databases was hampered by low bandwidth, the lack of a standardized interface, and platform dependencies (Duncan, 1995). The emergence of the Web helped fix the interface and platform problems, and high-speed connections are helping to fix the bandwidth problem. Yet another problem to be solved was how to access databases dynamically, avoiding the problem of having to convert database data manually to text suitable for transmission using the hypertext transfer protocol (HTTP). This problem has been solved through the use of programs and scripts that convert relational data to hypertext markup language (HTML) for transmission over the Web.

DATABASE SUPPORTING ROLES

Not every database application is directed toward public outreach and public access, although the public side often works in conjunction with the private side. This section briefly discusses types of databases and the support they provide to organizations.

Reference databases support retrieval of literature citations, abstracts, and hard data from vast databases. Most are made available on a subscription basis. They are seldom used for transaction processing but can be useful as part of the informational services of an organization. They are of vital interest to libraries, information agencies, and information research units within organizations, and they make up an important component in corporate knowledge-management operations, digital libraries, and data mining. They are available over the Internet from companies such as Lexis/Nexis, Ovid, Ebsco, Dialog, and Dow Jones.

Back-office databases are part of an internal information system supporting the functioning of the organization. Some portions of the databases may be made available to Internet users, but they are less often used in transaction processing. Typically, back-office applications use small or medium capability database management systems such as Access (Microsoft), Paradox (Corel Corporation), FileMaker (Filemaker), and MySQL (MySQL AB).

Front-office databases support organizational functions that reach out to customers, constituents, or clients. They support customer or public information, selling and buying transactions, data warehousing, business intelligence, customer service, banking, securities, and a host of commercial and noncommercial applications. Typical front-office databases include DB2 (IBM), Oracle (Oracle), SQL Server (Microsoft), Informix (IBM), and Sybase (Sybase).

DATABASES IN E-COMMERCE

E-commerce (EC) databases support interactive catalogs of products and services, customer profiles, ordering, paying, transaction history, billing, shipping, and inventory management (Turban, 2002). EC is only one environment delivering database information and conducting transactions over the Internet. Others include nonprofit organizations such as libraries, universities, service organizations, government agencies, clubs, and other organizations that have a need to provide information and data interactions to customers, the public, constituents, students, members, and the like. These noncommercial applications are no less important than EC, and the principles of database access over the Internet are the same.

Database-driven Web sites run the gamut from huge, technically complex systems running on many servers to small sites using one or two servers. In addition to the usual functions of shopping, billing, payment services, and information dissemination, some sites maintain customer relationship management services, personalization features, and statistical analysis. Many are not so complex. Regardless of size and complexity, they all face the same problems of limited bandwidth, problems delivering images, and slow database connectivity (Shand, 2000), but these problems are being addressed as the Internet evolves and database vendors add features to make their databases more Web-compliant.

DATABASE EVOLUTION

Database technology has gone through several definable periods from early file systems to object-oriented databases.

File Systems

Before and during the1960s, data processing was mostly an attempt to computerize manual file systems. These early file systems had the following characteristics: data was separated and isolated across the enterprise, it had to be duplicated for different applications using the same data, data structures were dependent on programs written to exploit the structures, there was incompatibility between applications, and most systems depended on fixed (programmed) queries with little flexibility (Connolly & Begg, 2002).

Along the way, people began to use the term "database" to describe those file systems, but by whatever name, they did not usually have the defining characteristics of a database: data independence, controlled redundancy, interconnectedness, security protection, and real-time accessibility (Martin, 1975, p. 25).

Most file systems were single entity, flat files in which a single file contained all the data needed for an application. Multientity databases were introduced in the 1970s with relationships among data maintained through a complex network of linked records. Models included *hierarchical databases* in which pointers addressed subordinate records and *network databases* in which the pointers went both ways in a many-to-many relationship. These linkages become quite complex, and programmers had to write complex code to navigate the links.

The 1980s brought "nonprocedural languages" for creating, maintaining, and accessing databases. These were written and sold without regard to their ultimate application. As these products matured, it became easier for people to create their own custom database management systems using wizards and simple program expressions, instead of writing long and complex procedural code. The data and programs to manipulate the data were becoming separate. These small- and medium-scale systems are still being used, most often in "back-office" environments, but with a significant presence in "front-office" environments as databases for Web access.

Relational Databases

Codd (1970) introduced the idea of relational databases. Unlike files in a hierarchical or network database, tables in a relational database are not physically connected. This provides a separation of the physical aspects of data storage from the logical aspects of data access. Rather than using internal pointers to relate data, relational databases use common fields to pull things together.

Relational database systems separate what the user sees from what the designer sees, and may even separate what the designer sees from what the network administrator sees. The data are independent from the applications that use them. A database consists of an application structure (user level) and a physical structure (physical level) with a schema or logical description as intermediary (the logical level).

The user level consists of "views" of the data to meet specific application needs. On the Web, the user level most likely consists of HTML forms that present descriptive text, images, menu items, buttons, text boxes and drop-down lists for requesting information, and so on. The user's view of the database is a series of choices and opportunities.

The middle, or logical level, consists of descriptions (attributes) of data, data entities, and relationships. The physical level deals with such things as tables, fields, joins, indexes, files, storage devices, server paths, and so on. This multitiered scheme allows data independence, limited redundancy, multiple sharing, and unanticipated retrieval.

Object-Oriented and Hybrid Databases

A complex, interactive, public-access environment may place demands on databases that the relational model cannot handle. For example, an EC or digital library environment might include structured data such as customer names and addresses as well as unstructured data such as text, images, videos, graphics, sounds, maps, and animation. The unstructured and loosely structured data are objects to be manipulated, and the evolving object-oriented approach extends database capabilities to include such complex objects and data types.

The relational model is too rigid to handle such a variety of data, but an object-oriented database (OODB) will provide the necessary flexibility (Abiteboul, 1997). An object may encapsulate into one package the data, relationships among the data, and methods for manipulating the data (Burleson, 1999). An object may contain a mix of text, multimedia, and methods appropriate to the object. Its data types may be defined specifically for the object, instead of relying on the limited set typically provided as part of a relational system. OODB is changing the face of databases and database access. Nonetheless, Burleson (1999, pp. 26–27) outlines problems to consider: OODB databases work best in dynamic, interactive environments but are unproved in large corporate environments. In addition, they are not completely accepted by major vendors, and there is a lack of standardization.

Possibly for these reasons, and certainly because of the large investment that organizations already have in relational databases, vendors have extended their traditional relational products into what are called object/relational, or hybrid, databases. Database systems, such as Oracle, Informix, Sybase, and IBM have added support for object features, including user-defined data types, rules, and functions (Burleson, 1999, p. 27). For example, Oracle8i includes three database "flavors": traditional relational, traditional relational extended to include object concepts and structures, and object-oriented based solely on object-oriented analysis and design (Loney & Koch, 2000). OODB is a complex technology, and, although it has strong adherents, its emergence has been slow. Relational databases will remain in service for many years.

INTERNET-ENABLED AND TRANSACTION DATABASES

Internet accessibility places greater demands on databases. Cox (2000) lists the following requirements of Internet-enabled databases:

- **Security:** The Internet is an open network; data must travel across many networks and network components. Steps must be taken to ensure the data will be kept secure.
- **Scalability:** Databases must be able to handle larger amounts of data and supply it to a larger number of users. Some database systems will not scale up for Internet use.
- **Support for Internet standards:** Databases must support an array of new standards, such as XML and Java.
- **Integration:** In an EC environment, public databases must work with other databases, legacy systems, third-party software, application servers (such as financial servers, banking), Web servers, messaging, and other middleware.

A typical EC data application will contain a mix of informational and transactional processes. Transactional databases support such things as ordering, bill paying, registration, inventory control, banking transactions, airline reservations, and stock trading. In the Internet environment, the demands placed on transaction databases are greater than those placed on informational or decision support databases because data changes more rapidly, sometimes involving near simultaneous changes to the data. The special requirements for transaction processing go by the acronym ACID, and these requirements apply to databases that support transactions:

- **Atomic:** There should be no incomplete transactions. For example, a sold item must both be added to a shipping order and removed from inventory. Furthermore, if any part of the transaction fails, the entire transaction should fail.
- **Consistent:** The output of the transaction should reflect the current data. Transactions cannot work with incomplete, out-of-date data, in keeping with the idea that transaction databases are usually more dynamic and rapidly changing than informational databases.
- **Isolation:** Transactions cannot reveal their results to other users, and they should not "collide." It is important that data being accessed are not, at the same time, being updated by another transaction. Most database systems provide some lockout method when two transactions occur nearly simultaneously. The system may lock out the field, record, or even the entire database, until the first transaction is complete.
- **Durable:** Changes made must be permanent, and there must be a way to recover from transaction failures. Either the transaction must be completed, or the database must be "rolled back" to a state before the transaction started. Log files are used to account for changes to the information, including information on when the change occurred.

Not every component of EC is transaction based, and not all EC activities demand ACID requirements. Advertising, browsing, and general information supply may not require ACID properties, whereas purchasing and payment processing may require them all (Umar, 1997).

KEYS AND RELATIONSHIPS

Relationships among data are expressed through the use of common fields among tables. Most often this is done through the use of keys. There are two types of keys. A *primary key* is a field or combination of fields that uniquely identifies each record—an item number in a sales catalog, for example. There can be no two products in the table with the same item number. A *foreign key* is a field in one table that is a primary key in another table. For example, a table about books may have a field showing the publisher's name. Another table, about publishers, will have a primary key field of publisher's name. The name is not the primary key in the book table, but it is a primary key in the publisher table. The linking of common fields

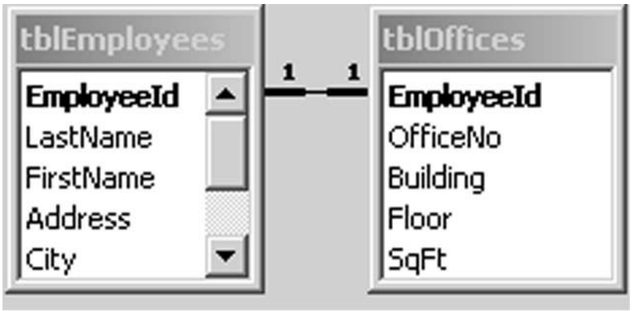


Figure 1: One-to-one entity.

among tables sets up relationships among the tables. The following relationships may exist.

One-to-None

This is not really a relationship. It is called a "single entity." All of the data are in a single table (in actuality, the database may consist of several tables, but they are all independent with no connections between them). The data has one "view": that which one application requires. Single entity databases are simple but inflexible. They work well for simple data management tasks but cannot meet the requirements of more sophisticated applications.

One-to-One

A record in one table relates to a maximum of one record in another table. This may be considered a single entity but with the data scattered across more than one table. For example, consider tables about managers and their offices. There would be a one-to-one relationship between *manager → office number*. This relationship is usually evident when a single-field primary key relates to a single-field primary key in another table, or when a foreign key in one table relates to a single-field primary key in another. Figure 1 illustrates the relationship. There may be good reasons to scatter the data across separate tables in a one-to-one relationship (security being one of the reasons), but often such data can be combined into a single table.

One-to-Many

A record in one table relates to zero, one, or many records in another table. This relationship is evident when a single-field primary key relates to part of a multiple-field primary key in another table. Figure 2 shows a customer table and an orders table. For any one customer there may be none, one, or many orders.

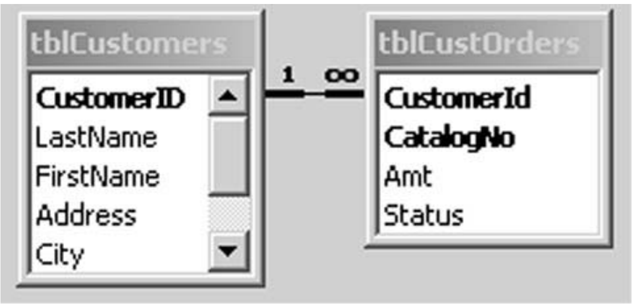


Figure 2: One-to-many relationship.

Many-to-Many

A record in one table relates to many records in another table,while, at the same time, a record in the second table relates to many records in the first table. An example is *customers ↔ products*. A customer may order many products and several customers may order the same product. This relationship usually requires setting up three tables with two one-to-many relationships. Figure 3 shows a table of customers, a table of products, and a table of customer orders. The relationships would be *customers ↠ orders* and *products ↠ orders*.

Recursive

A recursive relationship exists when a foreign key relates to the primary key *in the same table.* For example, a personnel table might include a field that shows the employee's supervisor. Both employee and supervisor records are in the same table because supervisors are also employees.

REFERENTIAL INTEGRITY

Database systems include a set of rules to ensure valid relationships between records in related tables. Specifically, they ensure that parent–child relationships match up. These rules, sometimes called "foreign key integrity rules," prevent adding a child record if there is no corresponding parent record, deleting a parent record when a child exists, and changing a parent's primary key if there are related child records.

An implementation of referential integrity rules may prevent such things from happening altogether, or it may allow "cascading" changes. For example, a cascading update automatically changes the matching value in all related records when the primary key of the parent record is changed. A cascading delete deletes related records when the parent is deleted.

NORMALIZATION

Codd (1970) introduced the idea of normalization when he introduced the idea of relational databases. These are practical rules of database design that we call "normal forms." Normalization has two functions: (a) to eliminate unwanted dependencies (see the discussion of Normal Forms), and (b) to reduce redundancy. A database designed in a helter-skelter fashion might work for a while, but problems will begin to crop up; simple additions, changes, and deletions become complex, requiring adding, changing, or deleting more data than intended. In addition, there may be unnecessary redundancy, adding complexity and wasting space.

Needless to say, these rules must be followed before a database can be made Web accessible. The slickest browser and the fastest server are of no value if the database is not organized properly. I briefly discuss the topic but leave the details for the reader to pursue in one of many textbooks on relational databases.

First Normal Form (1NF)

1NF states that the data must be cast into a table with one row for each record and one column for each field, with a primary key established that uniquely identifies

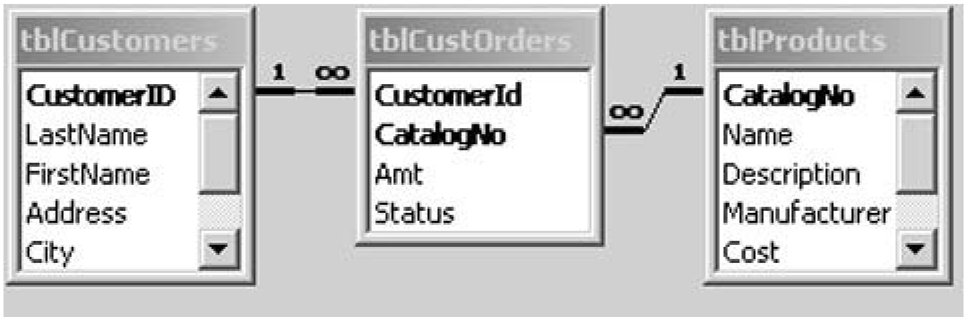


Figure 3: Many-to-many relationship.

each record. In addition, no field can have more than one value per record. A customer may have one last name, one address, and one telephone number, but the table that stores the customer information cannot also store more than one order for that customer.

To store the order information requires two tables: one for customers and one for orders. The customer table, with a primary key (PK) of CustomerID, contains information about customers but nothing about orders. Order data are stored in the order table with PK = CustomerID + OrderNo. The relationship, one-to-many, is made through the common CustomerID keys.

Second Normal Form (2NF)

A table satisfies 2NF when it satisfies 1NF and the entire primary key, not just part of it, identifies all fields. Violation of this rule, called partial dependency, occurs only with multiple-field keys. As an example, consider an order table that contains customers' names, addresses, and phone numbers along with ordered part name, part description, order status, and so on (PK = PartNo + CustomerID). This violates 2NF because the customer's address and telephone have nothing to do with PartNo, and the PartName field is dependent only on PartNo, not on CustomerID. It doesn't make sense to mix these fields, and doing so will cause update and deletion problems. The solution is to create three tables, one for Customers (PK = CustomerID), one for Parts (PK = PartNo), and one for Orders (PK = CustomerID + PartNo).

Third Normal Form (3NF)

A table meets 3NF when it is meets 1NF and 2NF and no field is dependent on another field that is not a key. This avoids a situation called transitive dependency. It is not always clear whether a particular design violates this rule, and it is easily (and sometimes deliberately) violated.

A typical example is when a table of catalog items has PK = CatalogNo, but the record also includes the manufacturer's address and telephone number. The manufacturer's name is correctly dependent on the primary key, but the address and telephone number are really dependent on the manufacturer's name. The solution? Remove the manufacturer's address and telephone number and put them in a separate table—a manufacturer's information table, keyed to the manufacturer's name. An important caveat: It may be worth violating this rule if you need to keep a "snapshot" of the conditions that existed at the time the item was put into the database. This is workable for archival files but not for active and volatile databases.

Fourth Normal Form (4NF)

2NF and 3NF have to do with the relationship between key fields and non–key fields. 4NF comes into play when parts of a key are dependent on other parts of the same key. This is called a "multivalued dependency" (Pascal, 2000). Because no part of a key may be blank, data must be entered in all parts of the key simultaneously, and all must be removed simultaneously. This is not always possible or desirable. The solution is complex and may require many related tables.

WEB DATABASE CONNECTIVITY

A Web browser usually should not connect directly to a database. The Web is based on HTML text files, but databases are structured and not easily handled directly by HTML. When it comes to moving data from the database to the user's browser, the alternatives are to convert database data into a series of text and HTML files and placing the files on the Web server (static access) or to run a Web page through a program that retrieves data from the database and creates HTML pages "on the fly" for return to the browser (dynamic access). These options are discussed in the sections that follow.

Static Database Access

Before 1995 the main strategy for providing relational database data over the Internet was to anticipate the kinds of queries users would want and then pre-create "views" of the data representing those queries. These queries were then marked up as HTML files and placed on a server. Pages created in this way represented a snapshot in time; they did not change as the database changed. They had to be re-created each time changes were made to the data. Worse, users could not search the database or choose particular items of data; they could only see what was previously saved in HTML format.

Databases with even minimal complexity are not easily converted into text files, and if they are, their dynamic characteristics are lost. The views available are dependent on what the database administrator decides to provide in the form of text files. The necessary step of creating text files from database tables is time-consuming and likely not done as often as it should be.

The development of HTML has been responsible for the success of the web. HTML is a markup language, however, and as such, it has little processing power. Web pages consisting only of HTML cannot search databases, make decisions based on conditions, do mathematical calculations, create animations, customize output for different users, respond to information entered on forms, and so on. These things require the addition of other languages to do the processing.

Dynamic Database Access

Dynamic database access means using Web pages to connect to databases in such a way that when users view the page, they are viewing the current state of the database. This method allows the user to connect to up-to-the-minute data, search it, and display it in different ways. This makes full use of an important characteristic of relational databases—the ability to reuse information in a variety of situations with multiple purposes.

I discuss several methods for creating dynamic database-driven Web pages, but first it is necessary to decide where the processing is to be done: on the user's computer (client-side) or on the server (server-side).

Client-Side Processing

Client-side processing is a method of having the server send back to the client a page containing both code and data. It is characterized by programs, or scripts, written in languages such as Java, JavaScript, or Visual Basic Scripting Edition (VBScript). The Web page calls programs on the user's computer to process the data. The user's computer does all the work; the server simply supplies the code and data.

Although this is less work for the server, there are problems with this approach:

- Not all browsers are guaranteed to support a particular language or its features.
- Debugging is difficult because the code must be designed for, and tested on, various browsers.
- The code is available to the user, a situation that may not be desirable in business environments.

Client-side processing is rarely used for database access, although there are a few systems that use it. One such is Microsoft Access, which includes a way to create what are called "Data Access Pages" (not to be confused with "Active Server Pages," discussed later). Database tables are converted into Extensible Markup Language (XML), which is sent to the user's browser for processing. The user must have the correct software to make use of it.

The use of XML to provide both data structure and data display shows promise. It provides many of the things found in databases: storage (XML documents), schemas (DTDs, XML schema languages), query languages (XQuery, XPath, XQL, XML-QL, QUILT, etc.), programming interfaces (SAX, DOM, JDOM), and so on. On the other hand, it lacks some of the things found in real databases: efficient storage, indexes, security, transactions with data integrity, multiuser access, triggers, queries across multiple documents, and so on. It is a powerful technology for many database applications, but Bourret (2002) argues that it may not work so well in production environments with many users, strict data integrity requirements, and the need for good performance.

Server-Side Processing

This method uses resources on the server, instead of the client, to process database data. After the server has retrieved and processed the data, it sends it back to the client as HTML for display on the browser. The server does all the work, and the client only needs to have a Web browser.

Server-side processing provides universal readability because the output is HTML. Browsers do not need special add-ons, and debugging is easier because it is not necessary to test and debug pages on the multitude of browsers that may be encountered. The server, not the client, works on the code and sends back the results, so the source code is not available to the user, an important consideration for EC (Kauffman, 1999).

Browsers connect to databases through a three-tier architecture (Watson, 2002, p. 452). The first tier is the browser, which sends Web page requests to the Web server. The second tier is the Web server that services the request and passes data request to an extension program. Extension programs are the workhorses of database access. They provide the connection to the database and translate Web data requests to database language. The third tier consists of the database server.

The second tier extension program accepts requests, converts them to a database-compliant form (such as OBDC SQL), and sends them to the database server (Lang & Chow, 1996). The database server processes the request against the database and sends the results back to the extension program, where it is converted to HTML and then sent to the Web server and browser.

Drivers

Because different databases organize data in different ways, the program to access a database must go through another program, called a driver, to give the database a more generic look. In essence, a driver has two faces (or interfaces). One face is proprietary and communicates with a specific database management system, and one face is standards-based and communicates with the application. Database management systems come with a set of drivers (sometimes called "native drivers") that interface with the particular database. Other drivers work with a variety of databases. For example, Open Database Connectivity (ODBC) is a standard developed by Microsoft but used by many vendors. Use of the standard makes it possible to access data from any database, regardless of which database management or query system is being used.

Another universal approach is Java Database Connectivity (JDBC) which was developed by Sun Microsystems. It allows Java programs to interact with any Structured Query Language (SQL)—compliant database. JDBC is designed specifically for Java programs. It is widely used but is not as language-independent as ODBC.

Identifying the Database

Web-based access requires some method of supplying information to allow the second tier to make a physical connection to the database. The necessary information might include the database name, the server path to the database, and driver name, user ID, and password. There are two methods of supplying the necessary information: Data Source Name (DSN) Connections and DSN-less Connections.

Data Source Name (DSN) Connections

In this approach, the database and driver are installed on the server, and the administrator assigns a DSN to the database. The DSN includes information about where the database is physically located, what kind of database it is, and what its name is. The Web page connects to the database by "calling" the DSN. Here is a typical DSN call to a database using VBScript:

```
Set conn = Server.CreateObject("ADODB.
  Connection")
conn.open "DSN=products"
```

The first line creates an object called "conn" and the second line opens the DSN identified as "products." Because the DSN is set up beforehand, the server knows to which database "products" refers and where it is located.

DSN-less Connection

This method does not require a DSN. Instead, the Web page contains all necessary information for the server to make a connection. Here is an example of a DSN-less call, again using VBScript:

```
Set conn = Server.CreateObject("ADODB.
  Connection")
conn.open "driver={microsoft access
  driver(*.mdb)}; dbq="& server.mappath
  ("serials")
```

Notice that this approach supplies the name of the driver, the type of database, the path to it, and the actual database name. All identifying information is contained within the call. This is a simple approach and does not require an administrator to set up a DSN.

Structured Query Language

SQL is a universal language to query, retrieve, add, and change database data. Most HTML-to-database connections make use of SQL. This chapter includes a few examples, but I do not attempt to describe the language here. The examples given use PHP as the underlying script (for more information about PHP, see the discussion about Embedded Programs).

SQL statements perform the following functions:

- SELECT receives a subset of data from one or more tables. It may be a subset of fields, a subset of records, or both. The following example selects a subset of fields and a subset of records from one table. The first line makes the connection, the second places the SQL statement in a variable, and the third executes the statement (subsequent examples will only show the middle line). This particular SQL statement retrieves the name, address, and telephone from the publisher table for those records where the city is Orlando:

```
$conn_id = odbc_connect('books','','') or
  die("Failure to connect");
$sqltext = ("Select pname, paddress, pphone
  from publisher where pcity = "Orlando"");
$result = odbc_exec($conn_id, $sqltext)
```

- JOINcombines data from more than one table, and there are several versions of this function: INNER JOIN combines data from two tables in which the keys match, LEFT JOIN combines all records from the "one" table and only matching records from the "many" table, RIGHT JOIN combines all from the "many" table and only matching records from the "one" table, THETA JOIN combines only records in which there is no common field match, and SELF JOIN combines a table with itself. Here is an example of an INNER JOIN:

```
$sqltext = ("SELECT m.manid, m.manaddress,
  p.cost FROM tblmanu INNER JOIN tblpart
  WHERE tblmanu.manid = tblpart.manid");
```

- This example retrieves the manufacturer's ID and address, from the many table and joins it with the cost from the part table for those records that have matching IDs.
- ACTION QUERIES are a class of queries that perform actions such as DELETE, INSERT, UPDATE. Here is a statement that decreases the price of all parts by 5%:

```
$sqltext = ("UPDATE tblparts SET price =
  price *.95");
```

The next issue is deciding where to put SQL statements. There are two approaches: Embedded SQL and Stored Procedures. The embedded SQL method embeds the SQL statement directly in the requesting Web page. Following is a search for publisher data in an ASP page, using VBScript (line numbers are given for reference; they are not used in actual pages):

```
1 <html> <head> <title>Publisher Name
  Search Response</title>
2 </head> <body>
3 <%
4 PName=request.form("PublisherName")
5 Set pub=server.createobject("ADODB.
  recordset")
6 sqltext="Select * from publisher  WHERE
  ppub like "&"'%" &  PName &  "%'" &  ";"
7 pub.open sqltext, "driver={microsoft
  access driver (*.mdb)}; dbq="&
  server.mappath("serials.mdb")
8 response.write "<p>The Publisher Data is
  as follows:<br>"
9 response.write "<p>Publisher Name: <b>" &
  pub("ppub") &  "</b><br>"
  ...other response.write statements...
10 %>
11 </body> </html>
```

Line 4 picks up the requested publisher's name from an HTML request form. Line 5 creates a record set object and names it "pub." Line 6 embeds the requested publisher's name in an SQL statement and stores the statement in a variable called sqltext. Line 7 opens the pub object and executes the SQL statement. An engine on the Web server (in this case, an ASP engine) sends the request to the database, which returns the correct publisher information. The rest of the page creates HTML statements out of the data and returns the page to the user's browser.

Queries based on SQL may be created ahead of time and stored as an object within the database using the stored procedures method. Instead of including the query in the requesting page, it asks the database system to run the stored procedure and return the resulting data for display. Here is portion of an ASP requesting page designed to run a stored procedure:

```
1 <html> <head> <Title>Items and
  Styles</title> </head>
2 <body>
3 <%
4 Set conn = Server.CreateObject("ADODB.
  Connection")
5 conn.open "driver={microsoft access
  driver (*.mdb)}; dbq="& server.mappath
  ("serials")
6 Set rs = Server.CreateObject("ADODB.
  Recordset")
7 rs.Open "qryitemstyles",conn
  ... other statements...
```

Line 7 opens and executes a stored procedure that has been created ahead of time to combine related data from two tables. The rest of the page takes the resulting data and creates HTML statements for sending to the browser.

TYPICAL SERVER-SIDE APPROACHES

Following are brief descriptions of typical approaches used to connect a Web browser to a live database. There are many other examples, each with variations, but they are similar to one of the approaches described here. They use different languages, such as Visual Basic, PHP, and Java to request user-entered information through a form, test for errors, display the results, and so on. Although they may be based on different *processing* languages, the common *retrieval* language is SQL.

These approaches divide into three methods of processing data: (a) using a separate program to do the processing (common gateway interface, or CGI), (b) using Application Program Interfaces (API), or (c) embedding the necessary programming language within the Web page and using a combination of CGI and a partner program (ASP, PHP, JSP, and so on).

Separate Programs: CGI

An early method of server-side processing, still widely used, is CGI. It is not a programming or scripting language but a standard against which to write programs for Web access. CGI programs may be written in almost any language that can be compiled, so long as the programmer follows the CGI standard.

Programs written according to the CGI standard take requests from a Web page, retrieve and process data from a database, and place the processed data into a file to be picked up by the user's browser (Khurana & Khurana, 1966).

One line of the requesting HTML page includes a call to a compiled CGI program residing on the server. Another line creates a Submit button. Following is an example (line numbers are for reference only):

```
1 <html> <head> </head> <body>
2 <form method="post" action="/cgi-bin/
  orders.exe/ksearch">
3 <p> Search Word: <input type="text"
  name="words"> [other html statements]
4 <input type="submit" value="ksearch">
5 </form> </body> </html>
```

Line 2 includes a call to a compiled CGI program called "orders.exe" residing in the cgi-bin folder of the server. When the user enters a keyword (line 3) and clicks the Submit button (line 4), the server executes the ksearch procedure of that program and returns a stream of text to be embedded in HTML for transmission and display on the user's browser.

CGI programs are platform-, software-, and database-independent, but they are not applicant-independent. Separate programs must be written, compiled, and loaded on the server for each application. CGI is widely used, and there are many free programs available for download.

It is slow because it must be loaded and run for each request. A more serious problem is that a CGI program might provide holes for mischief making. Hackers have used CGI programs to gain access to critical files on the server. As a minimum, any free, downloaded CGI programs must be examined for possible security problems, and the programs should reside on a server that contains no other critical files or information. Some network administrators simply will not allow CGI programs at all.

Application Program Interfaces

An API is a group of library routines that are called by the Web server as if they were part of its core program (Lang & Chow, 1996). The libraries are implemented as dynamic link library (DLL) modules or shared objects. Vendors can write programs that interact with the server operating system by calling these routines. The developer works with reusable blocks, making application development faster. DLLs serve much the same function as CGI programs, but instead of calling a separate CGI program, Web server software with an API extension intervenes and processes the request. This method is faster and takes up fewer server resources than CGI because APIs are part of the operating system. On the other hand, poorly written API extensions can bring down a server.

Embedded Programs

Another approach is to embed the programming language within the HTML page. In truth, these Web pages also interact with a “CGI-like” program, called an engine or a partner program, but those engines are part of, or added to, the operating system and are more general-purpose than CGI programs. The idea is to use a small CGI program partnered with a large server service that does much of the work (Lang & Chow, 1996). Application specifics reside in the Web page, and variations are brought about not by rewriting the engine but by writing different pages.

Typically, the browser requests the page containing the embedded script, either by URL or through the use of a menu or Web form. The server sends the page to the engine that reads the file and executes the script. The script contains calls to a database, and SQL statements seek out the data. The final product is an HTML page containing formatted data from the database.

All of the methods described in the following sections use this approach, and they are all similar in the way they work with databases. Of course, the devil is in the details.

Java Sever Pages

Developed by Sun Microsystems, Java Server Pages (JSP) use a combination of Java tags to retrieve data and HTML and XML tags to format and display it. A JSP engine receives requests from a client and generates responses for the client’s browser.

The JSP specification was developed through cooperation between e-commerce firms, vendors, and system designers. Tools were developed to integrate with and leverage existing expertise within the Java programming environment (Sun Microsystems, 2002). As a result of the cooperative effort, JSP will work on any Web or application server, any platform, and most any database product. The native scripting language is based on the widely used Java programming language, and the pages are compiled into what are called Java Servlets.

ColdFusion

ColdFusion was created by Allaire Corporation of Cambridge, Massachusetts, which later merged with Macromedia. Like most server-side products, it uses a CGI-like program partnered with an engine to access the database and build an HTML page incorporating both data content and data display. The resulting HTML is readable by any browser, but the server must be equipped with ColdFusion server software. The software can run on multiple platforms. It uses a scripting language called ColdFusion Markup Language (CFML).

PHP

PHP is a server-side scripting language that is becoming increasingly popular. It, too, embeds statements into HTML pages to build data content and data display. It derives its syntax largely from the C language and is cross-platform (although it is not supported by Macintosh). It is server and browser independent and provides native support for the most popular databases plus ODBC. It supports a large number of protocols such POP3, IMAP, and LDAP (Converse & Park, 2000).

Here is a sample program using PHP scripting language to connect to, and display data from, a Microsoft Access database though ODBC (line numbers are included for reference only):

```
1 <html> <head> <title>Publisher Name
  Search Response</title> </head>
2 <body>
3 <?php
4 $conn_id = odbc_connect('books','','')
  or die("Failure to connect");
5 $bibquery = ("Select * from publisher
  where ppub like '%$PublisherName%'");
6 $result = odbc_exec($conn_id, $bibquery)
  or die("Unable to search database");
7 $publisher = odbc_result($result, "ppub");
8 if (strlen($publisher) > 0)
9. {
10 $tel = odbc_result($result, "ptel");
11 $tel = "(".substr($tel,0,3).")
  ".substr($tel,3,3)."-".substr($tel,6,4);
12 print "<p>The Publisher Data is as
  follows:<br>";
13 print "<p>Publisher Name: <b>. "
  $publisher. "</b><br>";
  ... other statements...
?>
</body> </html>
```

This page has received a request from a Web form in which the requested publisher’s name is stored in PublisherName (line 5). The database call is to a DSN named “books” (line 4) and the SQL searches the table for PublisherName (lines 5–6). The rest of the program creates HTML out of the retrieved data.

Although a popular language, PHP is not widely used in the corporate setting. It is less object-oriented than some other languages, making it less suitable for large-scale, multiteam projects. It is a newer language and sometimes considered not mature enough for enterprise applications. Its heritage as a C language makes it easy for C programmers, but the language is not as intuitive or natural as other scripting languages. It is free and “open source,” which makes for dynamic evolution but at the cost of stability. For example, the online documentation includes “user comments” that describe problems (e.g., glitches, “won’t works,” and “work-arounds”), probably not something the enterprise data manager wishes to cope with.

Active Server Pages

Active Server Pages (ASP) are Microsoft’s method of providing server-side processing. Like the others, ASP imbeds scripting statements directly into the Web page. These statements may be written in Visual Basic Script or Jscript (Microsoft’s version of JavaScript).

ASP uses ActiveX Data Objects (ADO), a Microsoft-developed technology for working with databases on the Internet. ADO are prepackaged programming libraries that are built into Microsoft Web servers and called on to do some work when it needs to be done.

The example shown in the section on Stored Procedures is an ASP script. It uses a DSN-less connection to an Access database. See Yerkey (2001) for details about ASP technology. A demonstration and source code examples are also available (ASP Demo, 2001).

One serious limitation with ASP is that it works best in a complete Microsoft environment. The server must be running a Microsoft product. Combining Microsoft and other products is problematic. One way around this limitation is to use third-party software, such as Sun Microsystem's Chili!Soft ASP, which allows ASP pages to run under a variety of operating systems and platforms.

CONCLUSION

Many of the informational and transactional activities used by EC and other public-oriented organizations are supported by databases with the Web supplying the user interface and connections to them. Internet connections to dynamic databases place extra demands on the databases, and vendors are trying to satisfy those demands. Most of the connections use HTML over transmission control protocol/Internet protocol (TCP/IP) and rely on server-side processing. Systems use either CGI programs or an embedded mix of HTML and a scripting language to access databases and return information to the user.

The Web provides a global infrastructure, standards, and a presentation format, whereas database technology contributes storage techniques, query languages, structured data, and mechanisms for maintaining data integrity and consistency. Newcomers on the scene—object-oriented databases and semistructured data using XML—may change the rules of database design and provide more efficient access to complex data.

GLOSSARY

ACID An acronym to describe special requirements of computer transactions: atomic, consistent, isolated, and durable.

Application Program Interface (API) A group of routines called by a program to provide linkage to the operating system; APIs extend the capabilities of operating system software to allow a program to communicate with a database.

Common Gateway Interface (CGI) A standard for writing programs to access database data through Web pages.

Client-side processing A method of sending data and programs to the user's computer for processing.

Data source name The method of identifying and locating a database on a server.

Dynamic database access Using Web pages to connect to databases in such a way that when users view the page, they are viewing the current state of the database; The technology provides live, interactive access to databases, such that changes to the databases are immediately reflected in the Web page.

E-commerce In a narrow context, using computer systems and networks to conduct business. In a broader context, all transactional activities over the Internet.

Embedded programs Scripting language statements embedded in Web pages to access databases.

Internet-enabled databases Databases designed to meet the special demands of Web access.

Hybrid object/relational databases A traditional relational database that has been extended to include support of objects.

Keys Primary keys are fields in a table that uniquely identify each record; Foreign keys are fields in one table that are primary keys in another.

Normalization Rules for getting data ready for a relational database that are used to eliminate unwanted dependencies and to reduce redundancy.

Object-oriented databases A database system in which "data and instructions are combined into objects or modules that perform specific tasks when sent appropriate messages" (Watson, 2002, p. 581).

Referential integrity Rules for ensuring that parent–child relationships in databases remain synchronized.

Relational databases A collection of related data stored in multiple tables that are linked as required by comparing common columns.

Relationships Description of how data in tables relate to each other; Includes one-to-many, many-to-many, and recursive.

Server-side processing A method in which the processing is done on the server, which sends Hypertext Markup Language (HTML) statements back to the user's browser.

Structured Query Language (SQL) A universal language used to access and manipulate databases on a server.

Static access The method of providing preprocessed database data over the Web. The data is converted to HTML before use, and the resulting pages have no continuing relationship to the database.

Transaction A series of actions taken or operations processed as one unit of work in such a way that they must be entirely completed or aborted.

Transaction-based databases Databases used to support transactions such as ordering, bill paying, registration, inventory control, banking transactions, airline reservations, and stock trading.

CROSS REFERENCES

See *Active Server Pages; Common Gateway Interface (CGI) Scripts; Data Mining in E-commerce; Data Warehousing and Data Marts; Electronic Commerce and Electronic Business; Intelligent Agents; Machine Learning and Data Mining on the Web; On-Line Analytical Processing (OLAP); Personalization and Customization Technologies; Structured Query Language (SQL).*

REFERENCES

Abiteboul, S. (1997). Object database support for digital libraries. In C. Peters & C. Thanos (Eds.), *Research* *and advanced technology for digital libraries* (pp. 11–23). Berlin: Springer-Verlag.

Abiteboul, S., Buneman, P., & Suciu, D. (2000). *Data on the Web; from relations to semistructured data and XML.* San Francisco: Morgan Kaufmann.

ASP Demo (2001). Retrieved August 25, 2002, from http://www.informatics.buffalo.edu/faculty/yerkey/serials ASP/menu.htm

Bourret, R. (2002). *XML and databases*. Retrieved August 25, 2002, from http://www.rpbourret.com/xml/XMLAndDatabases.htm

Burleson, D. K. (1999). *Inside the database object model.* Boca Raton, FL: CRC Press.

Codd, E. F. (1970). A relational model for large shared data banks. *Communications of the ACM* 13, 377–387.

Connolly, T., & Begg, C. (2002). *Database systems* (3rd ed). Harlow England: Addison Wesley.

Converse, T., & Park, J. (2000). *PHP 4 bible.* Boston: IDG Books.

Cox, J. (2000). E-commerce changing the face of databases. *Network World, 17*(31), 38.

Duncan, R. (1995, August). Publishing databases on the World-Wide web. *PCTech,* 403.

Kauffman, J. (1999). *Beginning ASP databases*. Birmingham, UK: Wrox Press.

Khurana, G., & Khurana, B. (1996). *Web database construction kit.* Corte Madera, CA: Waite Group Press.

Lang, C., & Chow, J. (1996). *Database publishing on the Web and intranets.* Scottsdale, AZ: Coriolis Group Books.

Loney, K., & Koch, G. (2000). Oracle8i: The complete reference. Berkeley, CA: Osborne/McGraw-Hill.

Martin, J. (1975). *Computer data-base organization.* Englewood Cliffs, NJ: Prentice-Hall.

Pascal, F. (2000). *Practical issues in database management: A reference for the thinking practitioner.* Boston: Addison Wesley.

Shand, D. (2000, May 8). Simple sites, complex problems. *Computerworld,* 1980.

Sun Microsystems (2002). *Javaserver pages white paper.* Retrieved May 14, 2002, from http://java.sun.com/products/jsp/whitepaper.html

Turban, E. (2002). *Electronic commerce: A management perspective*. Englewood Cliffs, NJ: Prentice-Hall.

Umar, A. (1997). E-commerce bedrock. *Database Programming and Design, 10*(11), 33–41.

Watson, R. (2002). *Data management: Databases and organizations.* New York: Wiley.

Yerkey, A. N. (2001). Active server pages for dynamic database Web access. *Library Hi-Tech, 19,* 133–142.

Data Compression

Chang-Su Kim, *Seoul National University, Korea*
C.-C. Jay Kuo, *University of Southern California*

INTRODUCTION

Rapid advances in computing and communication technologies support lifestyles in the information age and the increasing demand for a wide range of multimedia services, including voice, audio, and video services. Multimedia data, however, require a vast amount of storage space or high transmission bandwidth. There are two approaches to support multimedia services. One increases the capacity of storage devices and transmission channels: the other compresses data compactly, thereby exploiting existing system capacity. The system capacity has increased continuously, but the need for more storage space and higher bandwidth has increased at an even faster pace. Moreover, new types of multimedia data are emerging. Therefore, extensive efforts have been made to develop data compression techniques to use limited storage space and bandwidth effectively.

Data compression is ubiquitous in daily life, and its application list is extensive. For example, it would take much longer to browse Web sites on the Internet that contain audiovisual information, if there were no data compression techniques. Data compression enables us to send photographs or music files efficiently via e-mail. Also, digital television broadcasting and the storage of high-quality video on digital video disk (DVD) are possible because of the video compression technology that efficiently compresses a huge amount of video data. Recently, mobile communication has become popular, and there is an explosive demand for mobile multimedia services, such as mobile Internet, m-commerce, and game services. Data compression is one of the key technologies that enables the transmission of multimedia data as well as voice over bandwidth–limited wireless channels.

Data compression can be classified into two categories (Witten, Moffat, & Bell, 1999): lossless and lossy techniques. In lossless compression, an encoder compresses the source data in such a way that the decoder can reconstruct the original data exactly from the compressed data. The lossless property is essential in some applications, such as the compression of texts, medical imaging, and satellite imaging for remote sensing, for which even small changes in data may lead to undesirable results. Also, lossless compression techniques are used to archive numerous types of data without loss of information.

In contrast, lossy compression attempts to achieve better compression by allowing loss of information during the compression procedure. There is a trade-off between the accuracy of the reconstructed data and the compression performance. Audio and video data are usually compressed in a lossy way, because the human perceptual system can tolerate a modest loss of audiovisual information. Moreover, audio and video data require a much larger storage space than text. For example, consider a video clip with a frame rate of 15 frames/s. If each frame consists of 176 × 144 pixels and the color information of each pixel is represented with 24 bits, the video clip requires the transmission bandwidth of about 9 Mbps. Suppose that we want to transmit the clip over 56-Kbps modem. The compression performance is often measured by the compression ratio, defined by

$$\frac{\text{Original data size}}{\text{Compressed data size}}.$$

In this case, we need a very high compression ratio, about 160 (= 9 Mbps/56 Kbps), thus it is reasonable to select lossy compression techniques.

In 1948, Shannon established the field of information theory (Shannon, 1948), which measures the amount of information in data quantitatively. Since the late 1960s, the subject of data compression has undergone extensive study, and numerous concepts and algorithms have been developed. In this chapter, we survey compression techniques and exemplify their application areas. It is organized as follows: First, we describe lossless and lossy compression techniques. Second, we present application areas of text compression, speech and audio compression, and image and video compression.

LOSSLESS COMPRESSION

In lossless compression, an exact duplicate of the original data can be reproduced after the compression and decompression procedures. We introduce several concepts and tools for lossless compression in this section.

Examples of Codes

In most digital computers, data are represented by binary numbers. Also, data are converted into binary numbers before the transmission over digital communication channels. Data compression or coding refers to the mechanism of assigning binary descriptions to samples of data source in a compact way. The set of binary descriptions is called a code, and each description is called a codeword.

Suppose that a communication system transmits a sequence of symbols (or samples) selected from an alphabet $\{A,B,C\}$, and that the probabilities of the symbols are given by

$$p(A) = 0.5, \quad p(B) = 0.25, \quad p(C) = 0.25. \tag{1}$$

Table 1 shows four example codes for this alphabet. The average length of a code is defined by

$$L = p(A)l(A) + p(B)l(B) + p(C)l(C),$$

where $l(A)$, $l(B)$, and $l(C)$ are the numbers of bits assigned to symbols A, B, and C, respectively. Code 1 seems to be most efficient because it requires the least number of bits to transmit each sample on the average, but it cannot be used in the communication system, because more than two sequences of samples can yield the same binary sequence. Let us assume that the decoder receives the binary sequence 10. It can be seen from Table 1 that both C and BA are encoded into the same sequence 10. Therefore, the decoder cannot determine whether it is generated by C or AB. To avoid this ambiguity, a code should be uniquely decodable, that is, every distinct sequence of samples should be mapped into a different binary sequence.

Code 2 is a uniquely decodable code. In Code 2, no codeword is a prefix of any other codeword. Thus, Code 2 is also called a prefix code. Figure 1(a) is the tree representation of Code 2. Each codeword is a leaf node, because Code 2 is a prefix code. Therefore, every binary sequence can be deciphered without ambiguity by traversing the tree. Let us decode 01110. Starting from the root node, the decoder traverses to the left child if the input bit is 0 and to the right child otherwise. When the decoder meets a leaf node, it outputs the corresponding symbol and starts again from the root node. Thus, 01110 can be deciphered into 0 11 10 = ACB. Note that every prefix code is a uniquely decodable code, and each codeword in a prefix code can be decoded instantaneously without reference to future codewords.

Table 1 Examples of Codes

Symbol	Probability	Code 1	Code 2	Code 3	Code 4
A	0.5	0	0	0	11
B	0.25	1	10	01	0
C	0.25	10	11	11	10
Average length		1.25	1.5	1.5	1.75

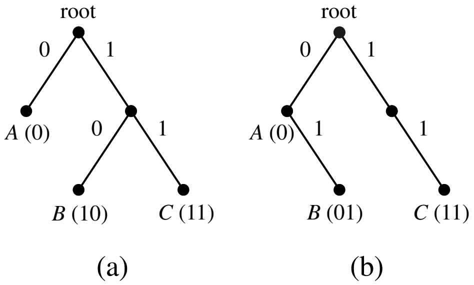


Figure 1: Tree representations of (a) Code 2 and (b) Code 3 in Table 1.

Not every uniquely decodable code is a prefix code, however. Code 3 is a uniquely decodable code, but not a prefix code. Figure 1(b) shows its tree representation. The codeword for A lies on the path from the root node to the codeword for B, because 0 is a prefix of 01. Therefore, when the decoder meets 0, it needs to check future bits to properly decode symbols. When 0 is followed by an even number of 1s, it is decoded into A. Otherwise, it is decoded into B. For example, 011110111 is decoded into 0 11 11 01 11 = $ACCBC$. In the worst case that a binary sequence is a concatenation of a single 0 and an infinite number of 1s, the Code 3 decoder need check infinitely many future bits to decode the first sample. The decoding of a nonprefix code hence may incur an unbounded delay and is more complex than that of a prefix code.

The goal of data compression is to find an optimal uniquely decodable code that provides the minimum average length. It can be shown (Cover & Thomas, 1991) that the minimum average length can be achieved also by a prefix code. Therefore, we can restrict our attention only to the set of prefix codes, which can be instantaneously decoded. Code 2 actually achieves the minimum average length. Code 4 is also a prefix code but less efficient than Code 2. It assigns a longer codeword to the symbol A than the symbol B, although A is more probable than B. It is readily seen that an efficient code should assign longer codewords to less probable symbols.

Entropy

In 1948, Shannon published "A Mathematical Theory of Communication," which founded the field of information theory. He developed a set of concepts that are necessary to understand the process of communication quantitatively. One of the concepts is entropy, which measures the amount of information or uncertainty in a data source. More specifically, the entropy of a data source is the number of bits required on the average to describe a sample of the data source. The entropy depends only on the probability distribution of the symbols if the symbols of the data source are independent and has nothing to do with what the data represent.

As mentioned previously, an efficient code should assign longer codewords to less probable symbols. Thus, the information or uncertainty of a symbol should be defined to be inversely proportional to its probability p. To find an adequate inversely proportional function, let us consider an alphabet that consists of 16 equiprobable symbols. Because the symbols have the same probability (1/16), they should be assigned the codewords of the same length. Obviously, the optimal code consists of the 4-bit binary representations of decimal numbers 0–15:

0000	0001	0010	0011	0100	0101	0110	0111
1000	1001	1010	1011	1100	1101	1110	1111

Hence, a symbol of probability 1/16 has 4-bit uncertainty. Similarly, considering an alphabet that consists of 2^n equiprobable symbols, it can be realized that a symbol of probability $1/2^n$ has n-bit uncertainty. In general, a symbol of probability p has $\log_2(1/p)$-bit uncertainty. Note that $\log_2(1/(1/2^n)) = \log_2 2^n = n$.

Consider a random data source X consisting of n symbols, which have probabilities $p_1, p_2, \ldots, p_n$. The entropy $H(X)$ of the data source is the average uncertainty of the symbols and thus given by

$$H(X) = \sum_{i=1}^{n} p_i \log_2 \frac{1}{p_i}. \tag{2}$$

This derivation of entropy is oversimplified. The entropy can be derived rigorously from three axioms (Shannon, 1948). Moreover, the entropy has the profound meaning that at least $H(X)$ bits/sample are required to encode a data source X losslessly, no matter what compression scheme is employed. In other words, $H(X)$ is the fundamental lower bound for the average length that a code can achieve. For example, the entropy of the alphabet in Table 1 is

$$0.5 \log_2 \frac{1}{0.5} + 0.25 \log_2 \frac{1}{0.25} + 0.25 \log_2 \frac{1}{0.25} = 1.5.$$

Therefore, no code can achieve less than 1.5 bits/sample on the average, which is the average length of Code 2 in Table 1.

Huffman Coding

Huffman coding constructs an optimal prefix code that provides the minimum average length (Huffman, 1952). It is based on the observation that the optimal code should allocate longer codewords to less probable symbols. Let us consider a random data source with n symbols. Without loss of generality, it is assumed that

$$p_1 \le p_2 \le \cdots \le p_{n-1} \le p_n,$$

where p_i denotes the probability of the ith symbol. Then, the codeword lengths should satisfy

$$l_1 \ge l_2 \ge \cdots \ge l_{n-1} \ge l_n,$$

where l_i is the codeword length of the ith symbol. Note that if $l_1 > l_2$, the last $(l_1 - l_2)$ bits of the longest codeword can be trimmed without violating the prefix condition. In other words, the code is not optimal. Hence, l_1 should be equal to l_2. Moreover, the two longest codewords should differ only in the last bit. If this is not satisfied, we can also trim the last bits of the two codewords without violating the prefix condition.

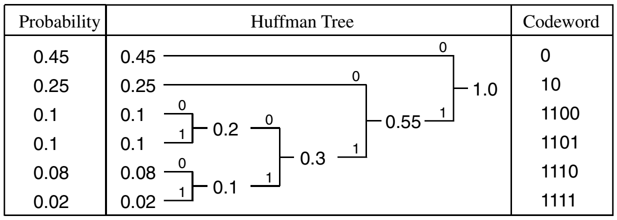


Figure 2: Construction of a Huffman code.

As shown in Figure 1(a), a prefix code can be represented by a binary tree. The above observations mean that the two least probable symbols should correspond to the two deepest nodes in the tree, and the two nodes should be siblings. Therefore, we can construct the optimal tree using a bottom-up method. Figure 2 illustrates how to construct a Huffman code. First, we merge the two least probable symbols while assigning 0 and 1 to the last bits of the corresponding codewords. The merged node is treated as a new symbol with the probability 0.1, which is the sum of the two child node probabilities. Second, we find again the two least probable symbols. There are three symbols with the probability 0.1 at this stage. We merge arbitrary two symbols among the three symbols into a new symbol and assign the probability 0.2 to the merged symbol. Proceeding in this way, we can merge all the nodes into the root node. Then, the original symbols are assigned the codewords that are the top-down concatenations of branch bits.

It can be shown (Cover & Thomas, 1991, Chapter 5) that the average length $L(X)$ of the Huffman code for a random data source X satisfies

$$H(X) \le L(X) < H(X) + 1. \tag{3}$$

For example, the entropy of the probability distribution in Figure 2 is 2.0872, whereas the Huffman code achieves the average length 2.15. $L(X) = H(X)$, if and only if the probability of each symbol is equal to 2^{-k} for some nonnegative integer k.

The upper bound in Equation 3 indicates that the Huffman coding scheme introduces at most 1 bit overhead compared with the entropy. This overhead may cause a problem in some situations. Suppose that a communication system transmits samples selected from an alphabet $\{A, B\}$, and $p(A) = 0.99$ and $p(B) = 0.01$. The entropy of the alphabet is 0.0808. As shown in the left column of Table 2, however, the Huffman code requires 1 bits/sample on the average, which is about 12 times higher than the entropy. We can reduce the overhead by encoding n consecutive samples simultaneously. In the middle column of Table 2, two consecutive samples are grouped into a block, and the Huffman code is constructed for the set of possible blocks. The Huffman code introduces at most 1-bit overhead per block, hence at most 0.5-bit overhead per sample. In general, if we design the Huffman code for

Table 2 Huffman Coding of Grouped Samples

No Grouping			Grouping of 2 Samples			Grouping of 3 Samples		
Symbol	Probability	Code	Block	Probability	Code	Block	Probability	Code
A	0.99	0	*AA*	0.9801	0	*AAA*	0.970299	0
B	0.01	1	*AB*	0.0099	10	*AAB*	0.009801	100
			BA	0.0099	110	*ABA*	0.009801	101
			BB	0.0001	111	*ABB*	0.000099	11100
						BAA	0.009801	110
						BAB	0.000099	11101
						BBA	0.000099	11110
						BBB	0.000001	11111
Average bits/sample = 1			Average bits/block = 1.0299 Average bits/sample = 0.515			Average bits/block = 1.06 Average bits/sample = 0.3533		

n sample blocks, it incurs only $1/n$-bit overhead per sample. Therefore, by increasing n, we can achieve an average bits/sample arbitrarily close to the entropy, which is the theoretical lower bound.

Arithmetic Coding

Huffman coding may require the grouping of infinitely many samples to achieve an average bits/sample equal to the entropy. The grouping of samples introduces transmission delay, however, and requires a higher computational load and a larger storage space for the codeword table. Arithmetic coding is an alternative coding scheme that provides a near-optimal compression performance without requiring the grouping of samples. It treats a whole sequence of samples, called a message, as a single unit but can be implemented to generate output codewords incrementally as samples arrive (Bell, Cleary, & Witten, 1990).

In arithmetic coding, a message is mapped to an interval of real numbers between 0 and 1. The interval length is proportional to the probability of the message and any number in the interval can be used to represent the message. Less probable messages yield longer sequences of bits, and more probable messages yield shorter sequences of bits.

For example, suppose that the alphabet is $\{A, B, C\}$, and the symbol probabilities are $p(A) = 0.6$, $p(A) = 0.3$, and $p(C) = 0.1$. As shown in Figure 3, the messages starting with A, B, and C are, respectively, mapped to the half-open intervals [0, 0.6), [0.6, 0.9), and [0.9,1), according to their probabilities. Let us encode a message BCA. Because it starts with B, it is first mapped to the interval [0.6, 0.9). Then, the interval is divided into three subintervals [0.6, 0.78), [0.78, 0.87), [0.87, 0.9), corresponding to the messages starting with BA, BB, BC, respectively. Note that the length ratio of the subintervals is set to

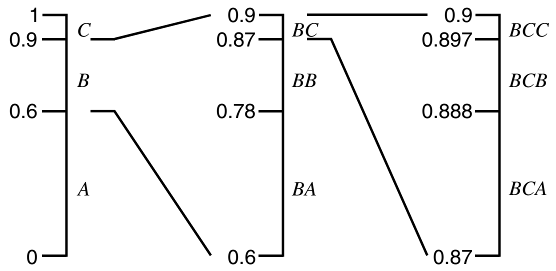


Figure 3: Illustration of arithmetic coding.

$$0.18 : 0.09 : 0.03 = 6 : 3 : 1 = p(A) : p(B) : p(C),$$

and 0.18, 0.09, 0.03 are the probabilities of the messages starting with BA, BB, BC, respectively. Similarly, [0.87, 0.9) is further divided into three subintervals, and the messages starting with BCA are mapped to [0.87, 0.888).

The two end points can be written in binary numbers as

$$0.87 = \frac{1}{2} + \frac{1}{2^2} + \frac{1}{2^4} + \frac{1}{2^5} + \cdots = 0.11011\cdots$$

and

$$0.888 = \frac{1}{2} + \frac{1}{2^2} + \frac{1}{2^3} + \frac{1}{2^7} + \cdots = 0.11100\cdots.$$

The prefix '0.' can be omitted, since every number within the unit interval [0, 1) is less than 1. The interval [0.87, 0.888) can be represented in the binary form [11011..., 11100...). The encoder can transmit an arbitrary number within this interval to specify that the message starts with BCA. For example, it can transmit the sequence of three bits, 111, which corresponds to 0.875 in decimal. Then, the decoder can follow the same division procedure in Figure 3 and know that 0.875 lies within the interval [0.87, 0.888), hence the message starts with BCA.

The decoder cannot, however, determine when the message ends; 0.875 also lies within the interval [0.87, 0.8808), which corresponds to the set of messages starting with $BCAA$. Thus, it can also specify the message $BCAA$. To make the arithmetic code uniquely decodable, we can add a special symbol to the alphabet, which indicates the termination of a message. If the special symbol is assigned a small probability and a message is long, the addition of the special symbol leads to only a minor degradation in the compression performance.

Arithmetic coding is superior to Huffman coding in many respects. It can achieve a near-optimal performance without grouping of samples. Also, it can be easily used in an adaptive way. In some applications, the probability distribution of alphabets cannot be predetermined and should be trained and updated based on the statistics of the data stream itself. To be adaptive to the change in the probability distribution, Huffman coding should regenerate the codeword table (Gallager, 1978), which is computationally inefficient. In contrast, arithmetic coding can adjust itself to the updated probability distribution simply by changing the length of intervals in the division procedure.

Dictionary Coding

Both Huffman coding and arithmetic coding implicitly assume that a data source generates a sequence of independent samples and do not consider the dependence between samples. Samples are dependent on one another in many data sources, however. For example, consider the following words from a text source:

Inte*net, bus*nes*, comp*ter, d*ta, compre*sion,

where * denotes a missing letter. We can reasonably recover the words to

Internet, business, computer, data, compression,

based on our knowledge of frequently used words. In other words, the missing letters can be retrieved by exploiting the dependence between letters, that is, by the context.

Dictionary coding achieves data compression by replacing blocks of consecutive samples with indices into a dictionary (Bell et al., 1990, Chapter 8). The dictionary should be designed to contain symbol blocks that are frequently used. If sufficient prior knowledge of a particular data source is unavailable, the dictionary can be built adaptively from the source itself. Many adaptive dictionary coding schemes are modifications of one of two approaches, LZ77 (Ziv & Lempel, 1977) and LZ78 (Ziv & Lempel, 1998). Popular file compression tools such as ZIP and ARJ are variations of LZ77. LZW is one of the most popular modifications of LZ78 and is used in the Unix command "compress," the V.42 bis standard for data compression over modem, and the graphics interchange format (GIF) for compressing graphical images.

Let us describe the encoding and decoding procedures of LZW. The encoder parses an input sequence into nonoverlapping symbol blocks in a dictionary and represents each block with its index to the dictionary. The blocks in the dictionary are also referred to as the words. The dictionary is initialized with the single symbols in the alphabet and grows during the encoding procedure. Suppose that the alphabet is $\{A, B\}$, and an input sequence is given by $ABABABBBBA$.

The encoder first parses the longest word in the dictionary from the input sequence. As shown in Table 3, the initial dictionary contains two words A and B. Thus, the longest word is A and is represented by the index 1. Then it is concatenated with the next sample B to form a new word AB in the dictionary. Similarly, the next longest word B is represented by the index 2 and is concatenated to the next sample A to form a new word BA. In this way, the input sequence can be parsed into

Table 3 Dictionary Construction in LZW Coding (input sequence = *ABABABBBBA*).

Index	Dictionary	Comment
1	A	initial
2	B	dictionary
3	AB	$(A) + B$
4	BA	$(B) + A$
5	ABA	$(AB) + A$
6	ABB	$(AB) + B$
7	BB	$(B) + B$
8	BBA	$(BB) + A$

$$(A)_{AB}\,(B)_{BA}\,(AB)_{ABA}\,(AB)_{ABB}\,(B)_{BB}\,(BB)_{BBA}\,(A),$$

where the parsed words are enclosed in parentheses and the newly formed words in the dictionary are written as subscripts. Consequently, the input is represented by the index sequence

$$1 \quad 2 \quad 3 \quad 3 \quad 2 \quad 7 \quad 1.$$

The decoder can reconstruct the input sequence from the index sequence by emulating the behavior of the encoder. It also starts with the initial dictionary, consisting of two words A and B. Thus, the first index 1 is decoded into A, which should be concatenated to the next sample to form a new word in the dictionary. The decoder looks ahead at the next index 2 that corresponds to a word B and inserts the new word AB into the dictionary. Similarly, the decoder reconstructs a word B from the second index 2, looks ahead at the next index 3 to find out that next sample is A, and inserts a new word BA into the dictionary.

Proceeding in this way, the third and fourth indices 3 are decoded into AB, respectively, and two new words ABA and ABB are inserted into the dictionary. Then, the decoder reconstructs the word B from the fifth index 2. The next step is to look ahead at the next index 7 to find out the next sample, denoted by *, and concatenate B and * to construct a new word B^* in the dictionary. This requires caution. Note that the index 7 cannot be directly parsed into a word, because the dictionary contains only six words and the seventh word is being constructed at this stage. We know, however, that the seventh word starts with B, thus the next sample * is B. Therefore, the seventh word in the dictionary is BB, and the decoder can continue.

Adaptive dictionary coding can be used to compress all types of data, thus being classified as a universal coding scheme. Care should be taken with its use, however. It can compress a data source effectively, provided that the source generates frequently recurring patterns or groups of samples. Text data provide a good example. In contrast, dictionary coding is not the most efficient way to compress natural image data, speech, and so on, which contain relatively few repetitive patterns. These data can be more effectively compressed by employing other approaches that are described later.

Run-Length Coding

Run-length coding (Jayant & Noll, 1984, Chapter 10) is suitable for compressing data that contain large segments of constant sample values. It replaces a sequence of samples by a sequence of pairs, $\{(s_k, r_k)\}$, representing sample values s_ks and run-lengths r_ks. For example, suppose that an input sequence is given by

$$AAAABBCCCBBBBBB.$$

It starts with a run of As with length 4, followed by a run of Bs with length 2. Thus, the first pair is $(A, 4)$ and the second pair is $(B, 2)$. In this way, the input sequence is converted into

$$(A, 4), (B, 2), (C, 3), (B, 6).$$

The sequence of pairs $\{(s_k, r_k)\}$ is usually encoded using the Huffman coding technique. Separate Huffman codeword tables can be constructed for the sample values s_ks and the run-lengths r_ks, or one table can be constructed jointly for the (s_k, r_k) pairs.

Run-length coding can compress binary and facsimile images effectively, which contain long runs of black (0) and white (1) pixels. Because white and black runs alternate, the colors (or sample values) of runs need not be encoded. To avoid ambiguity, the color of the first run is assumed to be white, and the white run of length 0 is encoded if the sequence starts with a black pixel. For example, a sequence of pixels

000111110111001

is converted into a sequence of run-lengths

0, 3, 5, 1, 3, 2, 1.

The decoder can know from the first run-length 0 that the sequence starts with a black pixel and interpret the following numbers as the lengths of black and white runs alternately.

LOSSY COMPRESSION

Lossless compression schemes can encode only data sources with discrete alphabets. Many data sources are analog however; that is, their symbols may have arbitrary values from a continuous range of amplitudes. The lossless description of a continuous sample requires an infinite number of bits. Thus, a finite representation of a continuous sample inevitably incurs some loss of precision, called quantization error. Moreover, even if a data source has a discrete alphabet, lossy compression techniques can describe a discrete sample more compactly by allowing some distortion.

In rate distortion theory (Cover & Thomas, 1991, Chapter 13), it was shown theoretically that there is a trade-off relation between the bit rate (in bits/sample) and the distortion. As more distortion is allowed, the data can be compressed more compactly. The goal of lossy compression techniques is to minimize the distortion subject to a given bit rate requirement or, equivalently, to achieve the lowest bit rate subject to a given distortion requirement.

A measure of distortion should be defined before describing lossy compression techniques. The most common measure is the squared error, defined by

$$d(x, \tilde{x}) = (x - \tilde{x})^2,$$

where x is the original sample and $\tilde{x}$ is the reconstructed sample. Also the mean squared error between two vectors $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\tilde{\mathbf{x}} = (\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n)$ is defined by

$$d(\mathbf{x}, \tilde{\mathbf{x}}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \tilde{x}_i)^2.$$

The squared error distortion is simple to use and has the intuitive meaning that it measures the power of the error signal. Thus, it is used in this section to evaluate the performances of lossy compression techniques.

Scalar Quantization

Quantization is the process of representing a large or infinite set of values with a smaller set (Gersho & Gray, 1991, Part II). In scalar quantization, the input and output sets are composed of scalar values. An N-point scalar quantizer partitions the real line into N cells $R_i = \{r : x_{i-1} < r \leq x_i\}$ for $i = 1, 2, \ldots, N$, and maps input values within a cell R_i to an output point y_i. Suppose that we have a data source modeled by a random variable X with probability density function (pdf) $f_X(x)$. Then, the quantizer yields the average distortion or the mean squared error given by

$$\begin{aligned} D &= \int_{-\infty}^{\infty}(x - Q(x))^2 f_X(x)\,dx \\ &= \sum_{i=1}^{N}\int_{R_i}(x - y_i)^2 f_X(x)\,dx, \end{aligned}$$

where Q denotes the quantizer mapping. Because the quantizer has N output points, each output point can be represented with $\lceil \log_2 N \rceil$ bits, where $\lceil a \rceil$ denotes the smallest integer greater than or equal to $\lceil a \rceil$. In other words, the bit rate of the quantizer is given by

$$R = \lceil \log_2 N \rceil \qquad \text{(bits/sample)},$$

when fixed-length codewords are used to encode the output indices.

For example, Figure 4(a) shows the input–output graph of an 8-point quantizer, whose mapping Q is given by

$$y = Q(x) = \begin{cases} 3.5\Delta & \text{if } x > 3\Delta, \\ 0.5(2n-1)\Delta & \text{if } (n-1)\Delta < x \leq n\Delta \ (n = -2, \\ & \quad -1, \ldots, 3), \\ -3.5\Delta & \text{if } x \leq -3\Delta. \end{cases}$$

Except for the two outer cells, every cell has the same length Δ. Thus, this quantizer is called a uniform quantizer. The bit rate is 3 bits/sample when fixed-length

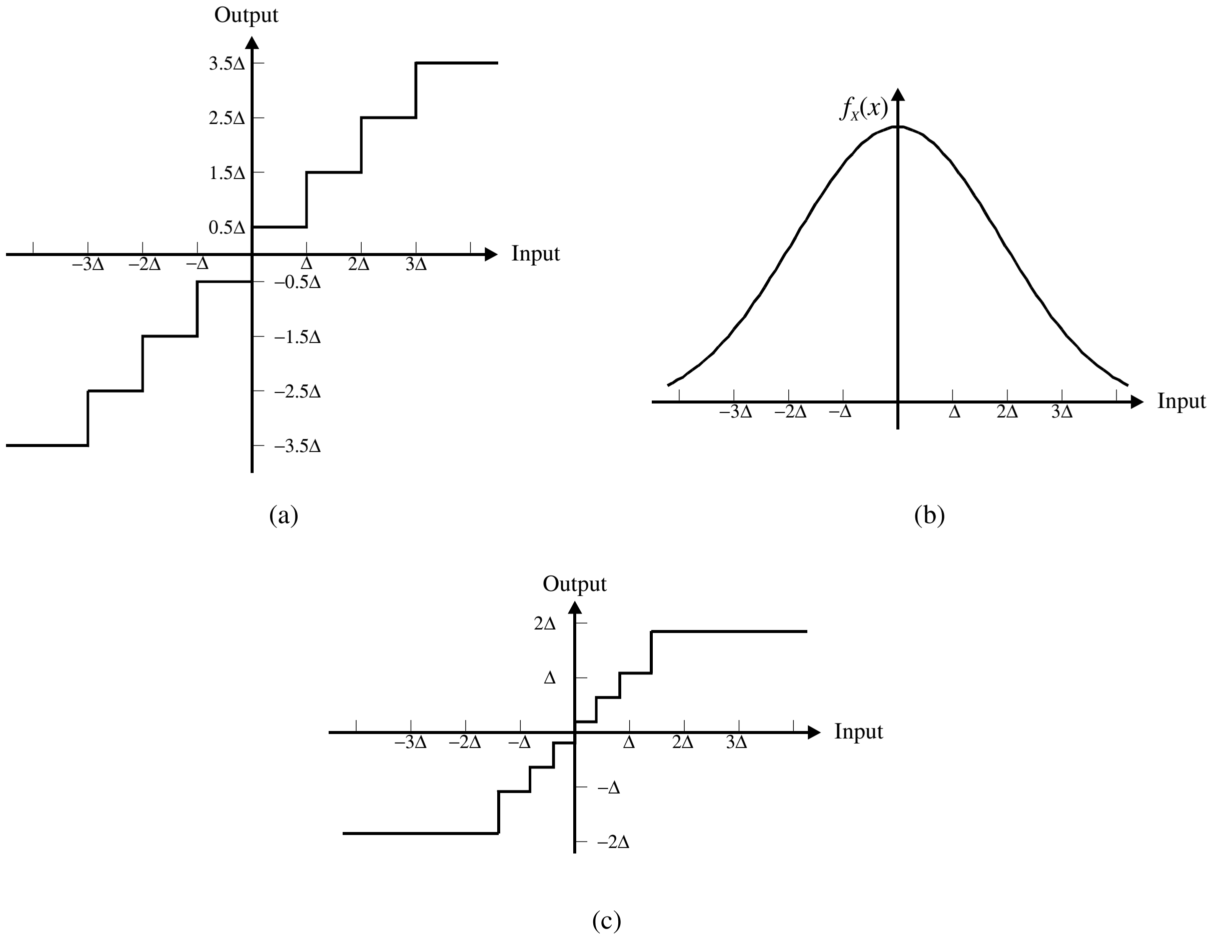


Figure 4: (a) A uniform quantizer, (b) a probability distribution function (pdf) of a random variable, and (c) Lloyd quantizer for the pdf in (b).

codewords are employed, because the quantizer has 8 output points. Output points can be described more compactly, however, because they do not have the same probability. Suppose that a random variable X has a probability distribution function (pdf) $f_X(x)$ shown in Figure 4(b). The probability P_i of the ith output point is given by

$$P_i = \int_{R_i} f_X(x)\,dx \quad (1 \le i \le 8).$$

Note that the fourth and fifth output points have higher probabilities than other points. Therefore, entropy coding schemes such as Huffman coding and arithmetic coding can achieve higher compression ratio than the fixed-length coding, by exploiting the nonuniform probability distribution of the output indices.

In practice, for a random variable with a pdf shown in Figure 4(b), the uniform quantizer is not the optimal 8-point quantizer that yields the lowest average distortion. Notice that a quantizer can reduce the average distortion by approximating the input more precisely in regions of higher probability. The optimal quantizer should make cell sizes smaller in regions of higher probability.

The Lloyd algorithm (Lloyd, 1982) is an iterative method that finds a locally optimal quantizer for a given pdf, which is based on two sufficient conditions for the optimal quantizer. First, an output point y_i should be the centroid of the corresponding cell R_i,

$$y_i = \frac{\int_{R_i} x f_X(x)\,dx}{\int_{R_i} f_X(x)\,dx}. \tag{4}$$

Second, a cell boundary point x_i should be the midpoint between two adjacent output points,

$$x_i = \frac{y_i + y_{i+1}}{2}. \tag{5}$$

The Lloyd algorithm first determines output points by Equation 4, after fixing boundary points. These output points are then used to select boundary points by Equation 5. These two steps are repeatedly applied until the change between the obtained average distortions becomes negligible. Because each step reduces the average distortion, the Lloyd algorithm always converges to a locally optimal quantizer. Figure 4(c) shows the Lloyd quantizer for the pdf in Figure 4(b). It can be observed that the Lloyd algorithm assigns smaller cell sizes around the origin to approximate regions of a higher probability more accurately.

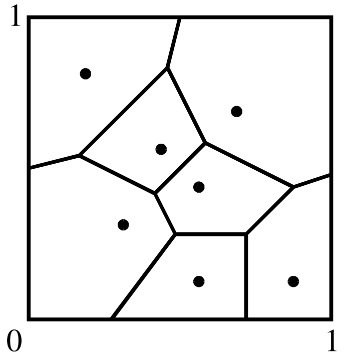


Figure 5: An example of a two-dimensional vector quantizer.

Vector Quantization

As shown in Table 2, the overhead of Huffman coding can be reduced by grouping samples together and encoding them as a single block. Moreover, as in dictionary coding, the grouping makes it easy for the encoder to exploit possible dependencies between samples. Vector quantization is the generalization of scalar quantization that quantizes a block of scalar samples jointly.

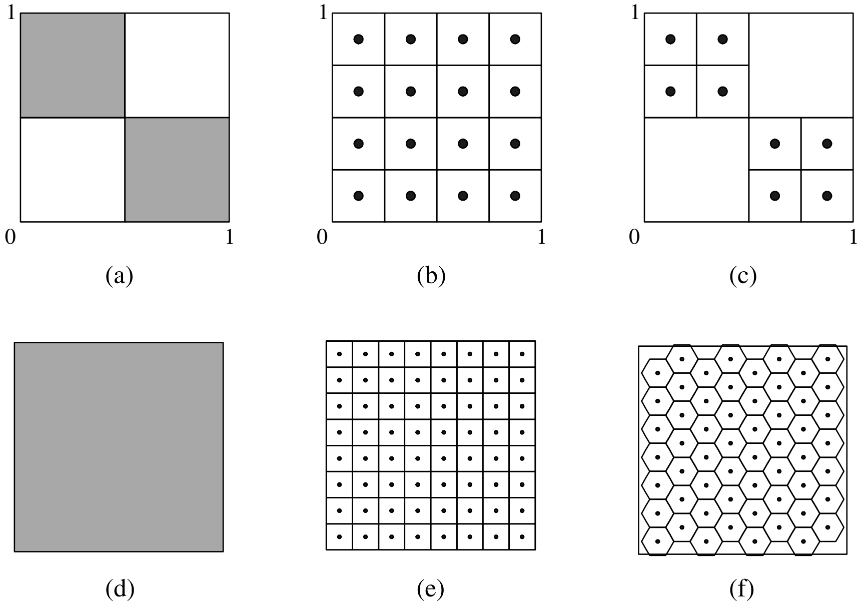
Figure 5 illustrates an example of two-dimensional vector quantizer, which partitions a unit square $[0, 1] \times [0, 1]$ into 7 cells. The black dot in a cell depicts the output point of the cell. In this quantizer, the configuration of these cells is completely determined by the locations of the output points. More specifically, the ith cell C_i is given by

$$C_i = \{x \in [0, 1] \times [0, 1] : d(\mathbf{x}, \mathbf{y}_i) \leq d(\mathbf{x}, \mathbf{y}_j) \text{ for } 1 \leq j \leq 7\},$$

where $\mathbf{y}_i$ denotes the ith output point. A vector $\mathbf{x}$ belongs to the ith cell if $\mathbf{y}_i$ is the nearest output point to $\mathbf{x}$ among all output points. This configuration of cells is called the Voronoi diagram of output points, and the quantizer is called the nearest neighbor quantizer. With the nearest neighbor quantizer, the encoder needs to store only the codebook consisting of output points, without requiring any information on the geometrical configuration of cells. For an input vector, the encoder compares its distances to the output points and records the index of the nearest output point.

If a set of output points is fixed, the Voronoi diagram is the best configuration of cells that minimizes the average distortion (Gersho & Gray, 1991, Chapter 11). Also, as in Equation 4, the output points should be centroids of the corresponding cells to minimize the average distortion. Therefore, in the same way as the Lloyd algorithm, we can design a locally optimal vector quantizer for a pdf by refining the configuration of cells and the locations of output points iteratively. This is called the generalized Lloyd algorithm (GLA).

Vector quantization has several advantages over scalar quantization (Gersho & Gray, 1991, Chapter 11). First, it can effectively exploit possible correlations between samples. Consider a two-dimensional random variable $\mathbf{X} = (X_1, X_2)$ that is uniformly distributed over the two shaded quadrants in Figure 6(a). Because the marginal pdf of X_1 or X_2 is uniform over the interval [0, 1], its optimal scalar quantizer is a uniform quantizer. If we encode each component independently by a 4-point uniform scalar quantizer, the unit square is partitioned into 16 rectangular cells as shown in Figure 6(b). This scheme requires 2 bits per component. However, note that components X_1 and X_2 are correlated with each other and the probability density over the upper right and lower left quadrants in Figure 6(a) is zero. Therefore, we can encode variable $\mathbf{X}$ with an 8-point vector quantizer in Figure 6(c),



(a) (b) (c)

(d) (e) (f)

Figure 6: Advantages of vector quantization over scalar quantization (Gersho & Gray, 1991, Chapter 11): (a) the probability density function, (b) the scalar quantization for (a), (c) the vector quantization for (a), (d) the probability density function, (e) a square cell partitioning for (d), (f) a hexagonal cell partitioning for (d).

which requires only 1.5 bits per component while achieving the same distortion as given in Figure 6(b).

As shown in Figure 6(b), scalar quantization of each component leads only to a rectangular cell partitioning in the two-dimensional space. Similarly, it leads only to a cubic cell partitioning in the three-dimensional space, and so on. In contrast, vector quantization has a freedom in choosing multidimensional cell shapes. Therefore, even if the components of a multidimensional variable are statistically independent, vector quantization can provide better performance than scalar quantization. For example, suppose that the components of a variable $\mathbf{X} = (X_1, X_2)$ are independent of each other and uniformly distributed as shown in Figure 6(d). Figures 6(e) and (f) are a square partitioning and a hexagonal cell partitioning for this pdf, respectively. The square cells and the hexagonal cells are of the same size. Thus, if we assume that the number of output points is very large and the boundary effects are negligible, the hexagonal partitioning requires the same rate as the square partitioning. Notice, however, that the hexagonal partitioning provides lower average distortion than the square partitioning, because the maximum distance between two points within the hexagon is smaller than that within the square.

Vector quantization is used to compress various kinds of data, including speech and image data. The main drawback of vector quantization is that it often requires a large storage space and high computational complexity. Consider a vector quantizer for image data, which encodes each 8×8 pixel region as a block at the rate of 0.5 bits per pixel. Each block is encoded with $32 = 8 \times 8 \times 0.5$ bits and the codebook is, hence, composed of $2^{32} \cong 4 \times 10^9$ vectors. If each pixel value is represented with a single byte, both the encoder and the decoder require about $8 \cdot 8 \cdot 4 \times 10^9$ bytes ($\cong$ 256 Gbytes) to store the codebook. Also, if the nearest neighbor rule is used to find the codeword, a brute-force encoder would compute 4×10^9 vector distances to encode each block, which is not acceptable in most applications. Hence, to reduce the storage space and the computational burden, several techniques have been developed such as lattice vector quantization, tree-structured vector quantization, and shape-gain vector quantization (Gersho & Gray, 1991, Chapter 12).

Predictive Coding

Many data sources generate a sequence of samples that are highly correlated to one another. Hence, in many cases we can predict the future sample based on the past and current samples, although perfect prediction is not necessarily possible. Prediction plays an important role in data compression systems because it enables the removal of redundancy in data to achieve a high compression ratio. In predictive coding, the encoder predicts each sample from the previous ones and transmits the prediction error instead of the original sample. Then, the decoder makes the same prediction as the encoder and uses the prediction error to compute the original sample. The entropy or energy of the prediction error is generally lower than that of the original sample; thus the prediction error can be more compactly compressed with lossless or lossy compression tools.

Let us consider an image compression system. Figure 7(a) shows the “Lena” image and its enlarged face

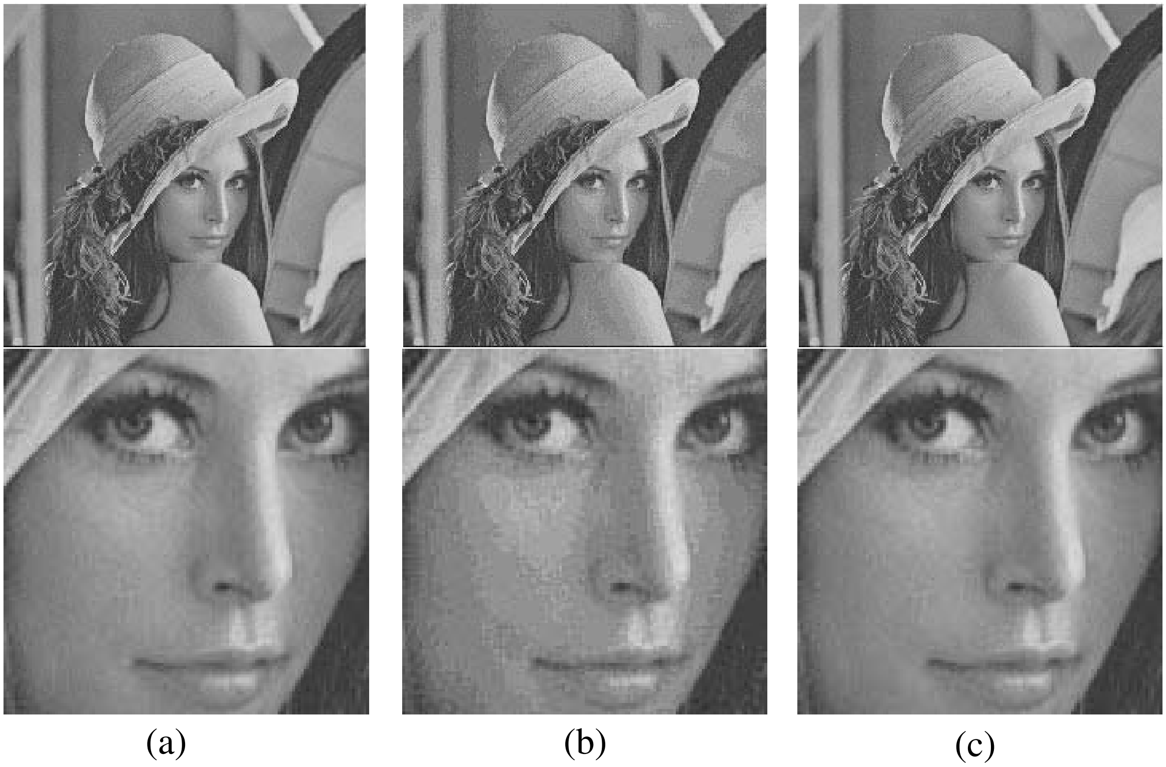


Figure 7: Lena images: (a) the original image, (b) the quantized image, and (c) the image obtained by predictive quantization.

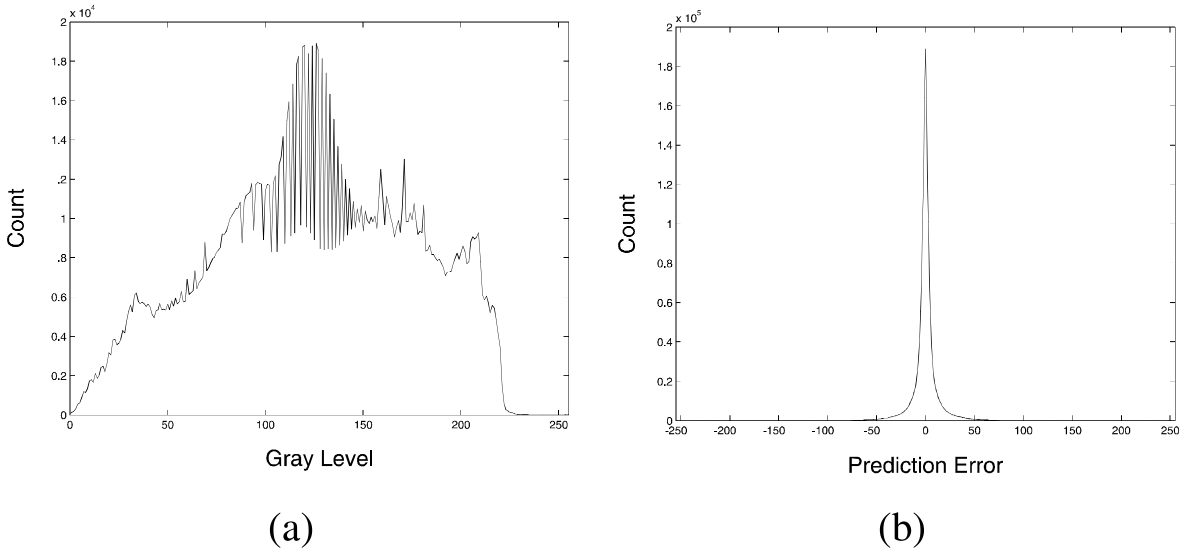


Figure 8: Histograms of (a) original gray levels and (b) prediction errors.

region. The image consists of 512 × 512 pixels. Each pixel has a gray level within [0,255] and is represented by a single byte. Figure 8(a) is the histogram of gray levels obtained from seven other test images. For each gray level, the histogram shows the number of pixels in the test images that have that gray level. Thus, after normalization, the histogram can be employed as an approximation of the pdf of gray levels. Based on the approximation of the pdf, we can design a 16-point scalar quantizer using the Lloyd algorithm. The 16 output points are

$$\{16,\ 30,\ 42,\ 58,\ 70,\ 84,\ 98,\ 114,\ 129,\ 143,\ 159,\ 173,\ 187,\ 200,\ 209,\ 217\}.$$

Figure 7(b) is the image quantized by this quantizer. It can be observed that the quantization error degrades the image quality. The rate is 4 bits per pixel, and the mean squared error is 18.974.

We can achieve better performance using predictive coding techniques, because adjacent pixels in typical images have similar gray levels. Suppose that the encoder compresses pixels in a raster scan order. Figure 8(b) is the histogram of prediction errors, when each pixel is predicted from the average of the past three pixels: left, upper, and upper right pixels. Note that the prediction errors are concentrated around 0 and much more compactly distributed than the original pixel values in Figure 8(a). For this distribution, the codebook of the 16-point Lloyd quantizer is given by

$$\{-56,\ -35,\ -25,\ -19,\ -15,\ -11,\ -7,\ -3,\ 0,\ 4,\ 8,\ 12,\ 17,\ 21,\ 32,\ 53\}.$$

Figure 7(c) is the image obtained by predictive quantization. During the prediction, the encoder uses the reconstructed values of past three pixels, instead of the original values, to avoid the mismatch between the encoder and the decoder. Then, the encoder quantizes the prediction error and transmits the output index to the decoder. In the decoder, the same prediction is performed, and then the quantized error is added to the prediction value to reconstruct a pixel. Note that Figure 7(c) provides better image quality than Figure 7(b). The mean squared error is 5.956.

The general structure of a predictive quantization system (Gersho & Gray, 1991, Chapter 7) is shown in Figure 9. This system is also called differential pulse code

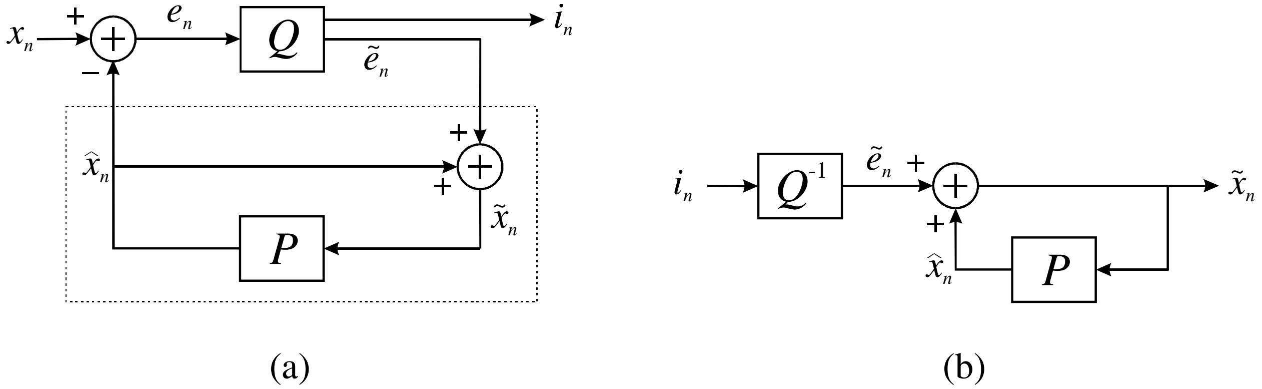


Figure 9: (a) The differential pulse code modulation (DPCM) encoder and (b) the DPCM decoder.

modulation (DPCM). The predictor P generates the prediction $\hat{x}_n$ of the nth sample x_n, based on the past reconstructed samples $\tilde{x}_{n-i}$ $(i = 1, 2, \ldots)$. The most common predictor is the linear predictor, given by

$$\hat{x}_n = \sum_{i=1}^{m} a_i \tilde{x}_{n-i}, \tag{6}$$

where a_i is a weighting coefficient. The quantizer Q quantizes the prediction error $e_n = x_n - \hat{x}_n$ into an output point $\tilde{e}_n$ and transmits the output index i_n. In the decoder, the dequantizer Q^{-1} converts the index i_n into the output point $\tilde{e}_n$ and adds it to the prediction $\hat{x}_n$ to reconstruct $\tilde{x}_n$.

A portion of the encoder, depicted as a dotted box in Figure 9(a), mimics the decoding operation, and uses the reconstructed samples in the prediction to avoid the mismatch between the encoder and the decoder. For example, consider a predictive quantization system, which uses the previous sample as the predictor and quantizes the prediction error to its nearest integer. Suppose that the encoder uses the original samples in the prediction and the input sequence is given by

$$0,\ 0.3,\ 0.6,\ 0.9,\ 1.2,\ 1.5,\ 1.8,\ 2.1,\ 2.4, \ldots.$$

Then, all the prediction errors are 0.3 and quantized to 0. Therefore, the decoder reconstructs the input sequence as an all-zero sequence, given by

$$0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0, \ldots.$$

In this case, the mismatch between the encoder and the decoder causes more and more errors in later samples due to error accumulation. On the other hand, if the encoder uses the reconstructed samples in the prediction, the decoder reconstructs the sequence as

$$0,\ 0,\ 1,\ 1,\ 1,\ 2,\ 2,\ 2,\ 2, .$$

and the mismatch problem is avoided.

Transform Coding

In transform coding, a sequence of source samples is represented by another sequence of transform coefficients, the energy of which is concentrated in relatively few coefficients. Because the coefficients have different statistics, they can be quantized more effectively than the original samples. In the decoder, the quantized coefficients are inverse transformed to reconstruct signal samples.

The discrete cosine transform (DCT) (Jayant & Noll, 1984, pp. 558–560) provides excellent energy compaction for highly correlated data. The DCT of a sequence $\mathbf{a}$ is related to the discrete Fourier transform (DFT) of the symmetric extension of $\mathbf{a}$ and can be computed by fast algorithms based on the fast Fourier transform (FFT). Thus, it is employed in several image and video compression standards. The DCT of an N-point sequence $\mathbf{a} = (a_0, a_1, \ldots, a_{N-1})$ is defined as

$$x_j = c_j \sum_{i=0}^{N-1} a_i \cos\left[\frac{\pi(2i+1)j}{2N}\right] \quad \text{for } j = 0, 1, \ldots, N-1,$$

where $c_0 = \sqrt{1/N}$ and $c_j = \sqrt{2/N}$ for $j = 1, 2, \ldots, N-1$.

The sequence of transform coefficients $\mathbf{x} = (x_0, x_1, \ldots, x_{N-1})$ can be inverse transformed into the original samples by

$$a_i = \sum_{j=0}^{N-1} c_j x_j \cos\left[\frac{\pi(2i+1)j}{2N}\right] \quad \text{for } i = 0, 1, \ldots, N-1.$$

This can be rewritten as

$$\mathbf{a} = \sum_{j=0}^{N-1} x_j \mathbf{b}_j,$$

where $\mathbf{b}_j$ is a basis vector, the ith element of which is equal to $c_j \cos[(\pi(2i+1)j)/(2N)]$.

Figure 10 shows the eight basis vectors for the 8-point DCT. When j is low, the samples in $\mathbf{b}_j$ are slowly varying, and the coefficient x_j represents a low-frequency

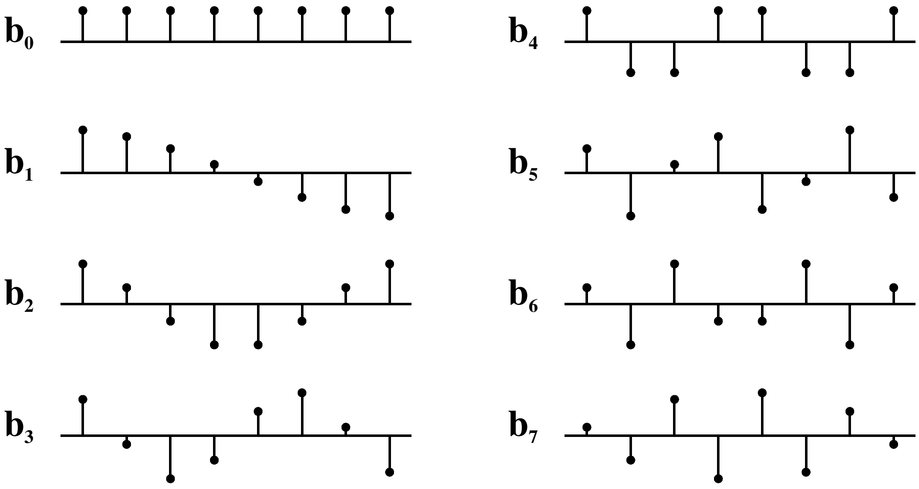


Figure 10: Basis vectors for the 8-point discrete cosine transform.

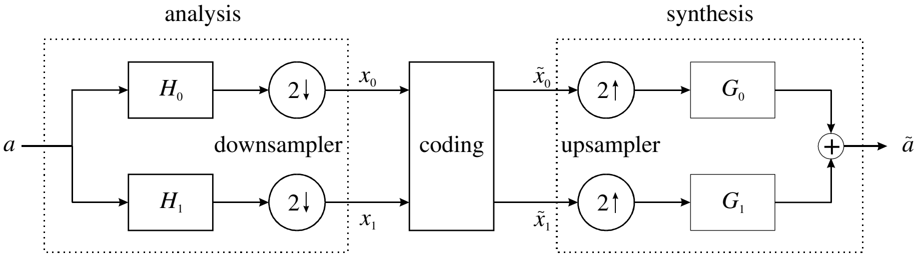


Figure 11: A subband coding system.

component in the signal. On the other hand, when j is high, x_j represents a high-frequency component corresponding to rapid variations between samples. As mentioned previously, many data sources generate a sequence of samples that are highly correlated. In such a case, high-frequency coefficients usually have smaller magnitudes than low-frequency coefficients. For example, consider the sequence of samples

$$(97,\ 102,\ 117,\ 135,\ 139,\ 140,\ 137,\ 130).$$

The 8-point DCT of these samples yields the coefficients

$$(352.5,\ -37.5,\ -25.2,\ 2.1,\ 1.8,\ 4.1,\ -0.7,\ -1.1).$$

Note that most energy is concentrated in the first three coefficients. Let us assume that the encoder transmits only the first three coefficients, after quantizing them into the nearest integers. Then, the decoder fills in the other coefficients with zeros and performs the inverse transform on

$$(353,\ -37,\ -25,\ 0,\ 0,\ 0,\ 0,\ 0)$$

to obtain approximate samples

$$(95.1,\ 104.6,\ 119.3,\ 132.7,\ 140.0,\ 139.9,\ 135.4,\ 131.4).$$

The approximation errors are relatively small, although only the first three quantized coefficients are employed in the reconstruction. In image compression systems, the two-dimensional DCT is employed, and the transform coefficients are encoded by more sophisticated coding schemes, such as zonal or threshold coding (Jain, 1989, pp. 508–510).

In most applications, the length of a signal is unbounded or too large to be transformed as a whole. Consequently, a source signal is partitioned into blocks of samples and the DCT is applied independently to the blocks. In this case, the DCT is called a block transform. The DFT, the DST, and the Hadamard transform are other examples of block transforms (Jain, 1989, Chapter 5). The main disadvantage of block transforms is that they cannot exploit possible correlations between blocks.

Subband and Wavelet Coding

Being different from block transforms, the subband transform involves no partitioning of samples, and thus possible correlations between blocks can be exploited. Figure 11 shows a 2-channel subband coding system, which transforms an input sequence into two frequency subbands. First, the encoder applies two analysis filters H_0 and H_1 to an input sequence, a, respectively. Then, the outputs are downsampled by a factor of 2 to generate two sequence x_0 and x_1. In general, H_0 is a low-pass filter and H_1 is a high-pass filter. Thus, x_0 represents low-frequency components in the original sequence, whereas x_1 represents high-frequency components. x_0 and x_1 can be quantized and coded differently according to their characteristics. In the decoder, the quantized sequences $\tilde{x}_0$ and $\tilde{x}_1$ are upsampled by inserting a zero between every two consecutive samples and then applied to synthesis filters G_0 and G_1 to reconstruct the input sequence approximately.

The analysis and synthesis filters can be designed so that the decoder reconstructs the original sequence losslessly, when there is no quantization error (i.e., $\tilde{x}_0 = x_0$ and $\tilde{x}_1 = x_1$). If this condition is satisfied, the set of the analysis and synthesis filters is called the perfect reconstruction filter bank. Several methods were proposed for the design of perfect reconstruction filter banks, driven by applications of subband coding of speech and image signals (Vetterli & Kovacevic, 1995, Chapter 3).

In the wavelet representation, a signal is hierarchically decomposed using a family of basis functions, which is built by scaling and translating a single prototype function (Mallat, 1989). The wavelet transform facilitates the multiresolution analysis of a signal, by reorganizing the signal as a set of details appearing at different resolutions. The wavelet transform has a close relationship with the subband transform, and the discrete wavelet transform (DWT) can be obtained by applying the two-channel subband transform to the low-pass subband recursively.

Figure 12 shows the subband decomposition of the "Barbara" image. First, we filter each row using two analysis filters and then downsample the filtered sequences to obtain low-(L) and high-(H) frequency subbands. In this figure, the 9-7 filter bank (Antonini, Barlaud, Mathieu, & Daubechies, 1992) is employed. Then, in the same way as the row decomposition, we decompose each column to obtain LL, LH, HL, and HH subbands.

This two-dimensional subband decomposition is recursively applied to the LL subband to obtain the DWT. Figure 13 shows the three-level wavelet decomposition of the "Barbara" image. The LL_3 subband approximates the original image at the coarsest resolution. The HL_i subbands contain vertical edge components at different resolutions ($i = 1, 2, 3$). Similarly, the LH_i and HH_i subbands

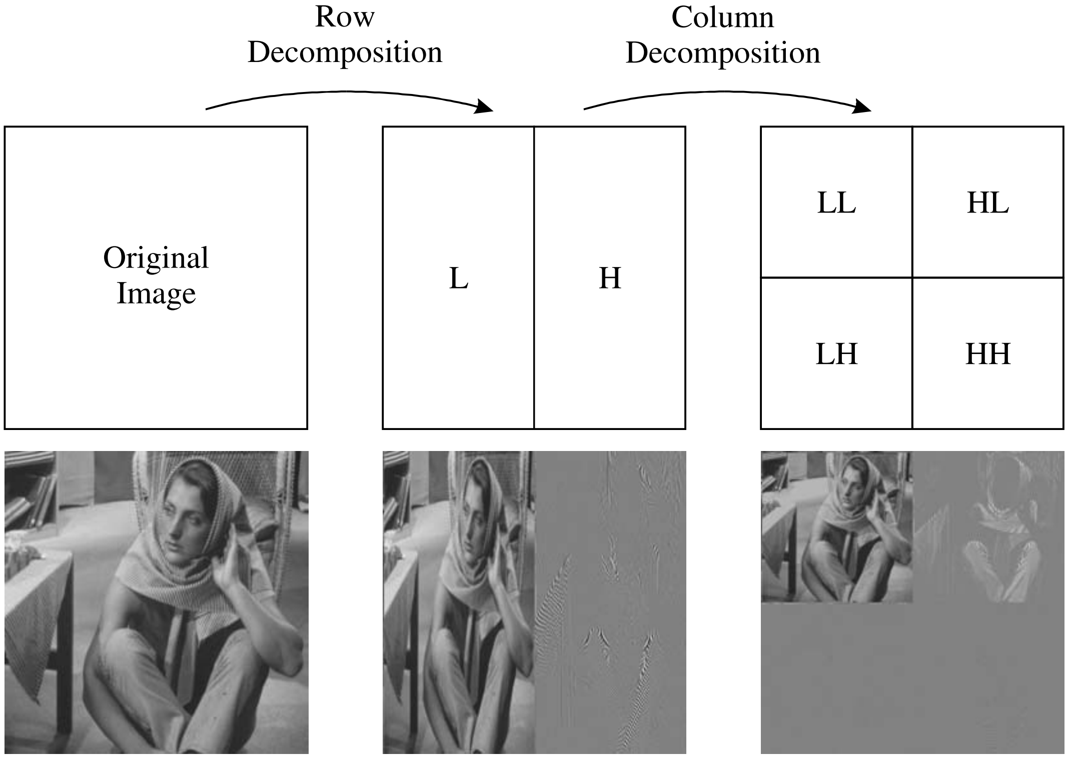


Figure 12: Subband decomposition.

contain horizontal and diagonal details, respectively. It can be seen that most signal energy is concentrated in the coefficients in the LL_3 subband. Therefore, the wavelet coefficients can be compressed effectively based on the adaptive quantization, by allocating a higher bit rate to the LL_3 subband than the other subbands (Vetterli & Kovacevic, 1995, Chapter 7).

Shapiro (1993) proposed an image compression algorithm, called embedded zerotree wavelet (EZW). His algorithm exploits interband correlations of DWT coefficients using zerotrees. Figure 13 illustrates a zerotree. The 21 coefficients, depicted by dots, represent the horizontal edge components of the same region in the original image. Thus, if the one coefficient in the LH_3 subband is less than a threshold, it is probable that the other 20 coefficients in the finer subbands also are less than the threshold. In such a case, the 21 coefficients are said to form a zerotree. Shapiro combined the concept of zerotrees with bit-plane coding and adaptive arithmetic coding to compress the DWT coefficients efficiently. The DWT is also used in the recent image compression standard JPEG2000 (Taubman & Marcellin, 2002).

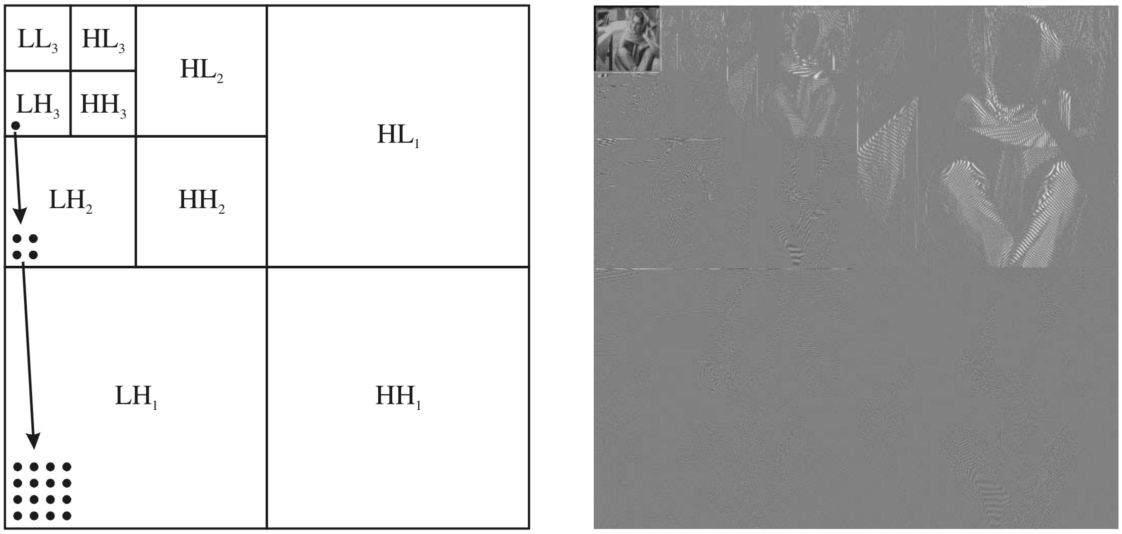


Figure 13: Three-level wavelet decomposition.

APPLICATIONS

A data compression system can be conceptually divided into a data modeler and a parameter coder. To compress a certain type of data, its characteristics should be first analyzed and modeled. Then, according to the data model, parameters are extracted from input data and compressed using one or more techniques in the previous two sections. The compression performance depends on how accurately the source data are modeled as well as how efficient the parameter coder is.

For example, letters in a text can be modeled as independent variables and each letter can be Huffman encoded, irrespective of its position in the text, according to the probability distribution of letters. Alternatively, we can adopt the model that the probability of a letter depends on the preceding letter, and design several Huffman codeword tables, one for each preceding letter. Similarly, the preceding *n* letters can be employed as a conditioning class, and a Huffman codeword table can be designed for each conditioning class. As *n* increases, the characteristics of text data can be more accurately approximated and text data can be more compactly encoded at the cost of higher complexities of the encoder and the decoder (Bell et al., 1990, Chapter 2).

Therefore, when designing the data modeler and the parameter coder, we need to consider several requirements of an individual application, including the desired compression ratio, tolerable distortion, and complexities of the encoder and the decoder. We describe main approaches and application areas of text compression, speech and audio compression, and image and video compression in the sections that follow.

Text Compression

By reducing the sizes of text files, we can use the limited storage space in a computer more effectively and save the communication time for sharing files between computers. In Unix environments, “compress” and “gzip” are often used as file compression tools. In personal computer environments, ARJ and ZIP are popular compression tools. These are all variations of Lempel-Ziv dictionary coding techniques. They can losslessly compress typical text data to 30–40% of its original size, because text data contain many redundancies and frequently recurring patterns.

Recently, context-based adaptive arithmetic coding has been studied to compress text data more efficiently (Witten, Neal, & Cleary, 1987). Moffat, Neal, and Witten (1998) reported that their arithmetic coder requires 2.20 bits/letter for a text file excerpted from the *Wall Street Journal*, whereas the gzip utility requires about 2.91 bits/letter. The text file has the zero-order letter entropy of 4.88 bits/letter, which is the minimum required bit rate when letters are modeled as independent variables.

Speech and Audio Compression

Speech and audio compression is used in many applications, including digital telephony, voicemail, voice recording, digital audio broadcasting, and digital audio storage. Speech and audio signals differ in dynamic ranges and listener expectations of offered qualities (Noll, 1997). Telephone speech has the frequency range of 300–3,400 Hz, and users are accustomed to tolerating some distortions. Telephone speech is often sampled with a sampling rate 8 KHz, and each sample is represented with 8 bits. Without compression, it requires the bit rate of 64 Kbps. On the other hand, high-quality audio has the frequency range of 20–22,000 Hz. On compact disks (CDs), audio signals are sampled at 44.1 KHz, and each sample is represented with 16 bits. Thus, each second of stereo music requires 176,400 (= $44{,}100 \times 2 \times 2$) bytes of data, corresponding to about 1.41 Mbps.

A popular speech coding technique is code-excited linear prediction (CELP) (Gersho, 1994). In CELP, the decoder generates the speech signal as the response of a synthesis filter to an excitation signal. The encoder periodically updates the synthesis filter by analyzing the speech waveform and determines the excitation signal for each segment of the waveform. The excitation signal is compressed by using vector quantization. A set of excitation vectors are stored in a codebook. For each time segment, the encoder transmits the optimal codeword to the decoder, which can generate the best approximation of the original waveform. The synthesis filter parameters are also quantized and transmitted to the decoder every 20 to 30 ms. CELP has been studied extensively, and many variations have emerged. It is the basis of many speech coding standards, such as G.723.1 and AMR. G.723.1 was standardized by International Telecommunication Union-Telecommunication Standardization Sector (ITU-T), and AMR by European Telecommunications Standards Institute (ETSI). G.723.1 supports 5.3 Kbps and 6.3 Kbps and is employed in video-conferencing applications. AMR supports eight bit rates between 4.75 Kbps and 12.2 Kbps and was adopted as the speech coder for the third-generation mobile communications by the 3rd Generation Partnership Project (3GPP).

Early work on audio compression is based on uniform or nonuniform scalar quantization of audio samples. However, these schemes cannot exploit statistical dependencies between samples, and thus cannot achieve a sufficiently high compression ratio. Most of recent audio compression algorithms rely on the subband or the block transform to decompose audio signals into several spectral components (Gersho, 1994). In human perception of audio signals, one sound can mask or block the perception of another sound. Based on the masking models, bits are allocated across the spectral components to reduce the required bit rate while maintaining the same perceptual quality. The Moving Pictures Expert Group (MPEG) has developed a series of audiovisual standards (Noll, 1997). MPEG-1/Audio consists of three audio coding schemes, called Layer 1, Layer 2, and Layer 3, which require bit rates between 32 Kbps and 448 Kbps. Among them, Layer 3 is the most involved and provides the best compression performance. It is also known as MP3 and popularly used for transmitting audio data over the Internet or storing music files in storage devices. MPEG-1/Audio has been extended to MPEG-2/Audio and MPEG-4/Audio, which provide better compression performance and support the coding of various kinds of audio data, such as the multichannel audio signal and synthetic speech.

Image and Video Compression

One of the earliest applications of image data compression is binary image coding for facsimile transmission. There are several compression standards for binary images, including MH, READ,MR, MMR, JBIG, and JBIG2 (Sayood, 2000, pp. 169–174). JBIG and JBIG2 provide better compression performance than other standards and are based on context-based adaptive arithmetic coders, called the QM coder and the MQ coder, respectively.

JPEG is a compression standard for gray-level and color images and is popularly used to transmit still images over the Internet or store photographs in digital cameras (Taubman & Marcellin, 2002, Chapter 19). It partitions an input image into blocks, consisting of 8 × 8 pixels, and transforms each block using the two-dimensional DCT. Then, each transform coefficient is quantized by a uniform quantizer. After the quantization, many coefficients are zeros because of the energy compaction property of the DCT. To encode consecutive zeros effectively, the quantized coefficients are zigzag scanned and run-length encoded with Huffman codewords. JPEG2000 is a recent compression standard for still images, which provides better performance than JPEG (Taubman and Marcellin, 2002, Part II). In JPEG2000, an input image is transformed into several subbands by using the DWT, and the bit-planes of subbands are transformed into a bit stream by an algorithm known as EBCOT and then encoded by the MQ coder. In addition to superior compression performance, it also supports efficient progressive compression, which is highly desirable when transmitting images over slow communication links. The decoder can first decode a coarse image and then gradually increase the image resolution or pixel accuracy as more data are received.

Compression technologies also offer the possibility of transmitting or storing a vast amount of video data in a compact way and are essential in many applications, such as digital television and high-definition television (HDTV) broadcasting, video on demand, videophone, and videoconferencing. Many efforts have been made in the video compression technology, and several international standards have emerged to support a wide range of applications (Sikora, 1997; Sullivan & Wiegand, 1998):

- ITU-T H.261: A video coder for videophone and videoconferencing applications. It was designed to compress video sequences, which consist mainly of head-and-shoulders of people. Its target bit rate range is 64–2,048 Kbps.
- ITU-T H.263: The first video coder designed specifically to support very low bit rate applications. Its target bit rate range was originally designed to be about 10–30 Kbps but has been broadened to 10–2,048 Kbps, because it provides better performance than the H.261 coder at any bit rate.
- MPEG-1: A video coder for storing movies on CD-ROM, covering bit rate range of 1–2 Mbps. It can provide VHS videotape quality at about 1.5 Mbps.
- MPEG-2: A video coder for digital TV/HDTV broadcasting and for storing video on DVD. It was designed to transmit high-quality, multichannel, multimedia signals at 2–15 Mbps.
- MPEG-4: A video coder similar to H.263, but it includes several functionalities to support flexible representation of visual data, such as object-oriented coding and three-dimensional mesh coding.

These standards compactly encode video sequences by exploiting temporal and spatial correlations via motion compensated prediction and DCT, respectively. More specifically, each frame is predicted blockwise from the previous frame using motion vectors. Then the residual frame consisting of prediction errors is encoded using a block DCT in a manner similar to JPEG.

CONCLUSION

Data compression is one of the key technologies that enables Internet and mobile communications. It has been extensively studied since the late 1960s, and various compression algorithms have been developed to store and transmit multimedia data efficiently. In this chapter, we classified data compression into lossless and lossy techniques and described key algorithms for each class. We then surveyed application areas of text compression, speech and audio compression, and image and video compression.

As new types of data emerge, researchers are considering modeling and compression techniques for them. For example, the compression of three-dimensional images is an important research topic, because three-dimensional images have the potential to be used in many applications, such as video gaming, virtual reality, and e-commerce. Progressive compression and error-resilient coding are also important issues, because they are essential in transmitting multimedia data over wireless channels that are bandwidth-limited and error-prone.

GLOSSARY

Code A set of binary descriptions assigned to samples of data source.

Codeword A binary description in a code.

Compression ratio The ratio of original data size to compressed data size.

Data coding A synonym for data compression.

Data compression A mechanism of representing data in a compact way.

Decompression A mechanism of reproducing original data from compressed data.

Decoder A device that reproduces original data from compressed data.

Encoder A device that compresses data.

Entropy The amount of information in a data source, or the minimum number of bits required on average to describe a sample of the data source.

Lossless compression Compressing data such that an exact duplicate of the original data can be reproduced at the decoder.

Lossy compression Data compression that allows loss of information.

CROSS REFERENCES

See *Speech and Audio Compression; Video Compression*.

REFERENCES

Antonini, M., Barlaud, M., Mathieu, P., & Daubechies, I. (1992). Image coding using wavelet transform. *IEEE Transactions on Image Processing*, *1*, 205–220.

Bell, T. C., Cleary, J. G., & Witten, I. H. (1990). *Text compression*. Englewood Cliffs, NJ: Prentice-Hall.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. New York: John Wiley & Sons.

Gallager, R. G. (1978). Variations on a theme by Huffman. *IEEE Transactions on Information Theory*, *24*, 668–674.

Gersho, A. (1994). Advances in speech and audio compression. *Proceedings of IEEE*, *82*, 900–918.

Gersho, A., & Gray, R. M. (1991). *Vector quantization and signal compression*. Norwell, MA: Kluwer Academic.

Huffman, D. A. (1952). A method for the construction of minimum-redundancy codes. *Proceedings of IRE*, *40*, 1098–1101.

Jain, A. K. (1989). *Fundamentals of digital image processing*. Englewood Cliffs, NJ: Prentice-Hall.

Jayant, N. S., & Noll, P. (1984). *Digital coding of waveforms—principles and applications to speech and video*. Englewood Cliffs, NJ: Prentice-Hall.

Lloyd, S. P. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, *28*, 129–137.

Mallat, S. G. (1989). A theory for multiresolution signal decomposition: The wavelet transform. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *11*, 674–693.

Moffat, A., Neal, R. M., & Witten, I. H. (1998). Arithmetic coding revisited. *ACM Transactions on Information Systems*, *16*, 256–294.

Noll, P. (1997). MPEG digital audio coding. *IEEE Signal Processing Magazine*, *14*, 59–81.

Sayood, K. (2000). *Introduction to Data Compression*. San Francisco: Morgan Kaufmann.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423, 623–656.

Shapiro, J. M. (1993). Embedded image coding using zerotrees of wavelet coefficients. *IEEE Transactions on Signal Processing*, *41*, 3445–3462.

Sikora, T. (1997). MPEG digital video-coding standards. *IEEE Signal Processing Magazine*, *14*, 82–100.

Sullivan, G. J., & Wiegand, T. (1998). Rate-distortion optimization for video compression. *IEEE Signal Processing Magazine*, *15*, 74–90.

Taubman, D. S., & Marcellin, M. W. (2002). *JPEG2000: Image compression fundamentals, standards and practice*. Norwell, MA: Kluwer Academic.

Vetterli, M., & Kovačević, J. (1995). *Wavelets and Subband Coding*. Englewood Cliffs, NJ: Prentice-Hall.

Witten, I. H., Moffat, A., & Bell, T. C. (1999). *Managing gigabytes: compressing and indexing documents and images*. San Francisco: Morgan Kaufmann.

Witten, I. H., Neal, R. M., & Cleary, J. G. (1987). Arithmetic coding for data compression. *Communications of the ACM*, *30*, 520–540.

Ziv, J., & Lempel, A. (1977). A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, *23*, 337–343.

Ziv, J., & Lempel, A. (1978). Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory*, *24*, 530–536.

Data Mining in E-commerce

Sviatoslav Braynov, *State University of New York at Buffalo*

INTRODUCTION

Data mining in e-commerce is a rapidly growing field, the development of which is driven by the explosive growth of online information and the economic need to provide personalized products and services. Although data mining applications have been around for several decades, e-commerce provides new challenges and new opportunities for the field. First, e-commerce data differs significantly from standard offline data in that it is distributed, highly volatile, unstructured, and heterogeneous. Second, e-commerce creates better opportunities for studying customers' behavior and providing highly personalized content.

This chapter is intended to help business managers and application developers in understanding the emerging field of mining e-commerce data, and attempts to accomplish the following:

- To summarize the driving forces for using data mining applications in e-commerce
- To provide an overview of existing technologies that can be used for mining e-commerce data
- To survey novel data mining applications, such as recommender systems, systems for log file analysis, Web marketing systems, and systems for customer profiling
- To summarize the open problems and current trends in data mining applications in e-commerce

DATA MINING IN E-COMMERCE

Basic Definitions

Data mining tools and applications perform data analysis and reveal interesting data patterns contributing to business strategies, market research, inventory management, and so forth. In a broad sense, the term "data mining" usually refers to the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data (Fayyad, Piatesky-Shapiro, Smyth, P., & Uthurusamy, 1996). There are several other terms with similar meaning, such as knowledge mining, knowledge discovery in databases, pattern analysis, and so on. Data mining is an interactive and iterative process consisting of several steps:

- Specification of the problem: Developing an understanding of the domain knowledge and identifying the goal of the data mining processes.
- Data collection and preparation: Removing noisy and inconsistent data, handling missing data fields, dimensionality reduction, transformation into forms appropriate for mining, integration of multiple data sources, and so on
- Analysis of data with data mining techniques: Applying a particular data mining method to extract data patterns
- Evaluation of the results according to previously established measures: Separating truly interesting patterns from uninteresting, spurious, and irrelevant patterns
- Interpretation and visualization of the results
- Making decisions or acting on the results

The main steps in a data mining process are shown in Figure 1. The data mining process may be iterative and contain repetitions and loops between any two steps. Data mining usually aims at achieving one of two goals: (a) verification and validation of previously discovered patterns and (b) the discovery of new patterns. When new patterns are discovered, they can be divided into two groups: descriptive and predictive. Descriptive patterns explain the general characteristics of data, whereas predictive

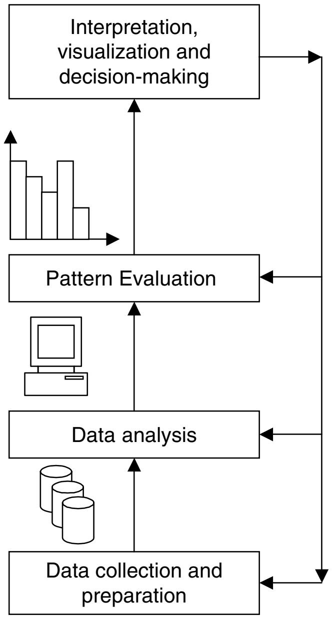

Interpretation, visualization and decision-making
Pattern Evaluation
Data analysis
Data collection and preparation

Figure 1: The main steps in the process of data mining.

patterns help make predictions about future trends, find missing or unavailable data, and so on. Predictive patterns are usually discovered by building models based on a set of training data. The success of predictive models usually depends on the training data used to generate the model.

Driving Factors for Data Mining in E-commerce

E-commerce Data

E-Commerce is an ideal domain for data mining. Compared with traditional offline business, e-commerce provides *large quantities of data* that keep growing at exponential speed. The enormous amounts of data generated from a Web site, for example, could provide meaningful insights into what type of visitors are likely to make purchases, what are the most typical navigation patterns, and what pages are most/least visited. The fast-growing e-commerce data exceeds human abilities for data collection, analysis, and decision making. Manual data analysis tools, such as statistical packages and query tools, are extremely time-consuming and fail to provide important data patterns. What makes e-commerce especially attractive for data mining is the ability to both generate and collect data automatically. Online transactions provide the opportunity to design, implement, and employ systems specialized in collecting data for the purposes of data mining (Kohavi & Provost, 2001). E-commerce systems can collect data that is otherwise difficult or costly to access and are able to keep track of customers' actions in virtual stores, that is, what articles they look at, how they browse a catalog, what items they put into their shopping cart. In contrast, tracking costumers in brick-and-mortar stores with in-store videos is difficult and costly.

Quality of data is another feature that makes e-commerce especially suitable for data mining. Collecting data electronically could significantly reduce the noise associated with manual data entry and processing. Noisy and corrupt data can hide some important patterns, introduce inconsistencies, or lead to misleading interpretations. Many online data collection systems provide validation functionality that helps correct erroneous data. For example, many Web-based hypertext markup language (HTML) forms automatically validate the user's input and check for inconsistencies.

Another factor facilitating data mining in e-commerce is *the measurable return on investment* (Kohavi & Provost, 2001). The Web allows for direct detection, collection, and analysis of revenue-related events such as Web site hits, clickstreams, purchasing events, and so on. Data mining results could be directly transformed into numeric measures of the success of Web design and the efficiency of marketing campaigns. In addition, data mining results are directly actionable by changing the Web sites design, improving inventory and category management, or targeting particular customers via Internet marketing. After interpreting the results of data mining, business analysts could pass recommendations to the developers of the online store, who will make necessary changes to improve, for example, the store's banner efforts, product assortment, checkout process, and product layout.

Another important advantage of e-commerce data mining is the possibility of conducting controlled experiments and measuring their effect on business goals. Direct marketing, for example, can target particular control groups, and the results can be data mined for predictive behavior modeling of consumers.

Mass Personalization

Personalization is an important driving factor behind the success of data mining applications in e-commerce. Personalization is the use of technology and customer information to tailor e-commerce interactions between a business and each individual customer (http://www.personalization.org). In general, personalization refers to making a Web site more responsive to the users' individual needs. This includes providing individually tailored products, services, and information. The goal of personalization is to serve customers better by anticipating their needs, customizing services and products, and establishing a long-term relationship encouraging customers to return for subsequent visits. Personalization is playing a constantly increasing role in e-commerce, in which customer experience is crucial to building customer loyalty to an online store. Personalization is usually based on building predictive models of customer behavior, preferences, and interests. Given its ability to build successful predictive models, data mining is an excellent personalization approach for building customer profiles, providing recommendations to the customers, and delivering personalized Web content. Most of the existing personalization tools make extensive use of different data mining techniques.

OVERVIEW OF BASIC DATA MINING TASKS AND TECHNIQUES

Data mining is an interdisciplinary field with a large variety of methods and algorithms developed in different research areas such as artificial intelligence, pattern recognition, machine learning, statistics, and databases. Its primary focus is on data representation, data storage and access, efficiency and scalability of algorithms, human–machine interaction, visualization, and so on. Data mining techniques may have different goals depending on the intended results of the data mining process. Most common data mining tasks include the following:

- **Classification:** the task of mapping a data item into one of several predefined classes. Each class is associated with a class label attribute, and the process of classification can be viewed as assigning labels to data items. In some cases, classification could be implemented as either supervised or unsupervised learning. In supervised learning, a training set is used in which the class label of each data item is known beforehand. The training set is analyzed by a classification algorithm, and the algorithm is then applied to a test data set to estimate the accuracy of the classification. Whereas supervised learning uses a training set in which the class label of each data item is known, in unsupervised learning the class labels and the number of classes may not be known in advance. Common applications of classification include, for example, loan payment prediction, customer credit policy analysis, and customer segmentation for target marketing.
- **Regression:** the analysis of the dependency of some variable (called the dependent variable) upon the values of other variables (called the independent variables). Regression could be linear (with linear dependency among variables) or nonlinear. Regression can be viewed as a special type of prediction in which one could predict the value of the dependent variable using the known values of the independent variables. Regression has multiple applications in e-commerce. For example, it can be used to predict the consumer demand for a new product as a function of advertising expenditures.
- **Clustering:** the task of grouping data items into classes based on the similarity between items. Clustering is a typical descriptive task in which the data is described using a finite set of categories or clusters. Clustering is used in marketing, for instance, to segment customers based on their purchasing patterns and demographic characteristics.
- **Dependency analysis:** the task of finding a model that describes dependencies between variables or attributes. Dependencies can be described at the structural or quantitative level (or both). The structural level specifies the direction of dependencies, that is, which variables are dependent on other variables. The quantitative level, on the other hand, specifies the strength of dependency between variables. Dependency analysis is used in numerous applications including learning propensity to purchase, real estate appraisal, fraud detection, and so on.
- **Change and deviation detection:** the task of discovering changes in data from previously measured values, or discovering data items that are significantly different from the remaining set of data. Deviation detection is used to detect unusual usage of credit cards or telecommunication services. It is also used in customized marketing for identifying groups of customers with extremely low or extremely high incomes.

The rest of this section provides a brief overview of the most common techniques used in data mining. Because of the wide variety of techniques the overview is by no means exhaustive.

Association Rule Mining

Association rule mining looks for interesting relationships among items in a given data set. It explains which attributes or which items tend to appear together. For example, the information that customers who buy coffee and sugar also tend to buy milk could be represented by the following association rule:

$$\text{buy}\,(X, \text{coffee}) \text{ and buy}\,(X, \text{sugar}) \Rightarrow \text{buy}\,(X, \text{milk})$$

In general, an association rule is an implication of the following form:

$$A \Rightarrow B,$$

where A and B are sets of items. Several items constitute a transaction. A transaction is defined as a set of items bought together.

The intended meaning of this association rule is that consumers who buy all items in itemset A also tend to buy all items in itemset B. Because thousands of association rules can be generated from large transaction databases, some weak and nonsignificant associations have to be filtered out. To eliminate spurious associations, two measures of rule interestingness are used: minimum support and minimum confidence. The minimum support measure ensures that the co-occurrence of itemsets A and B is frequent enough to signal a strong correlation between all items. In the example, a minimum support of 10% means that in at least 10% of all transactions, consumers have bought coffee, sugar, and milk. In other words, a minimum support of x% means that at least x% of transactions contain all items in A and B. The minimum confidence measure ensures that there is a strong relation between itemset A and itemset B. That is, the occurrence of itemset A implies the occurrence of itemset B with high probability. A minimum confidence of 60% means that at least 60% of consumers who purchased coffee and sugar also bought milk. In other words, a minimum confidence of x% means that at least x% of transactions containing all items in A also contain all items in B.

Association rule mining can be divided into two phases: (a) mining frequent itemsets, that is, itemsets occurring at least as frequently as some predefined threshold, called minimum support count, and (b) generating association rules from frequent itemsets. Usually only strong association rules are generated, that is, association rules satisfying minimum support and minimum confidence.

One of the most popular algorithms for mining frequent itemsets is Apriori (Agrawal & Srikant, 1994). As the name suggests, Apriori uses prior knowledge to reduce the size of the search space. Another efficient algorithm for association rule mining is the frequent pattern growth

(FP-growth), which uses divide-and-conquer strategy and compresses the database representing frequent items into a frequent-pattern tree (Han, Pei, & Yin, 2000).

Association rule mining can be used for dependency analysis, classification, and prediction. Common applications include market basket analysis, customer segmentation, and fraud detection.

Bayesian Belief Networks

A Bayesian belief network (BBN) is a directed acyclic graph (i.e., a graph with pointed arcs and no cycles) in which the nodes represent random variables and arcs represent probabilistic dependency. Each BBN incorporates two types of knowledge: structural knowledge, specifying which variables are dependent on each other, and quantitative knowledge specifying the strength of the dependencies. Structural knowledge is captured by the topology of a BNN, whereas quantitative knowledge is captured by conditional probability distributions describing the relationship between a node and its parents. One or more nodes in a network could be selected as output nodes representing particular classes. This makes BBN useful for probabilistic classification, for which the result is a probability distribution over classes.

A BBN is usually trained using a set of training examples. Two components of a BBN can be learned: graphical structure and conditional probabilities. The training is based on a theorem proved by Thomas Bayes, an 18th-century clergyman. The theorem specifies how conditional probabilities are updated as new data are observed. There are a variety of learning algorithms for BBN, depending on whether the structure is fixed or unknown and whether variables are hidden or observable. Bayesian belief networks are typically used for dependency analysis and classification.

Decision Trees and Rule Induction

A decision tree is a special type of tree where each nonterminal node represents a test on an attribute of a data item, each branch represents an outcome of the test, and terminal nodes (leaves) represent classes or class distributions. To classify a particular data item, one starts at the root node and tests the attribute values of the item against the decision tree by tracing the tree from the root to a terminal node. When a leaf is reached, the data item is classified as one belonging to the class associated with the leaf. An example of a decision tree indicating whether a consumer is likely to buy an insurance policy is shown in Figure 2. Each leaf node represents a class: buys or doesn't buy. Nonleaf nodes represent test on an attribute: age or income.

A decision tree can easily be converted to a set of if–then classification rules. A single rule is created for each path from the root to a leaf node. Each test along the path corresponds to a conjunction in the left-hand side (the *if* part) of the rule. For example, the decision tree in Figure 2 can be converted to the following set of rules:

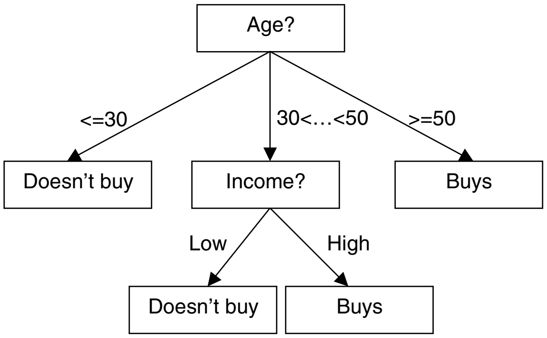


Figure 2: An example of a decision tree.

IF Age < = "30" THEN buys_insurance = "No"

IF Age > "30" AND Age < "50" AND Income = "Low" THEN buys_insurance = "No"

IF Age > "30" AND Age < "50" AND Income = "High" THEN buys_insurance = "Yes"

IF Age > = "50" THEN buys_insurance = "Yes"

One of the most widely used algorithms for decision tree induction is ID3. It uses an information-theoretic measure of the quantity of information, called entropy, defined in the late 1940s by Claude Shannon. The algorithm starts at the root node and incrementally builds the tree by selecting one attribute test for each node. From among all remaining attributes, the algorithm selects the attribute with the greatest entropy reduction. In this way, the algorithm minimizes the information needed to classify data items and the expected number of tests.

Artificial Neural Networks

An artificial neural network is a parallel and dynamic system consisting of many interconnected input/output units, designed to model the behavior of biological neurons. Each unit has many inputs that combine into a single output value. Units are connected, so that the outputs from some units are used as inputs to other units, and each connection has a weight associated with it. All inputs to a single unit are combined as a weighted sum, and the resulting input value is transferred to an output value via an activation function. The most common activation functions are step activation function, linear function, logistic sigmoid function, and hyperbolic tangent function.

An artificial neural network can learn a model of the training set that can subsequently be used for classification or prediction. The goal of the training is to adjust the weights so that the output of the network is as close to the desired output as possible, for as many examples of the training set as possible. The most common learning technique, back-propagation, was developed by John Hopfield. Back-propagation calculates the output for every example from the training set, finds the error by taking the difference between the calculated result and the expected result, and feeds the error back through the network. During the back-propagation phase, the weights are adjusted to minimize the error.

As with its biological counterpart, every unit in an artificial neural network has the property that small changes in inputs can have large effects on the output and vice versa. The property that small changes matter is usually called nonlinear behavior. What makes neural networks especially attractive is their ability to model this nonlinear behavior, their robustness, and their fault tolerance. When they degrade, they tend to degrade gracefully, making the reduction in performance less obvious. In addition, they are tolerant to noise in the input data. On the other hand, neural networks have been often criticized for their poor interpretability. It is difficult for a human expert to associate any meaning with the learned weights.

Clustering Techniques

Clustering is the process of grouping data into clusters, based on the principle of maximizing intraclass similarity and minimizing interclass similarity. In other words, objects in one cluster are similar to one another and are dissimilar to objects from other clusters. A typical application of cluster analysis is discovering homogeneous subpopulation of customers in marketing databases and characterizing each population based on purchasing patterns.

There are a variety of clustering algorithms. A good review of clustering algorithms is provided by Jain, Murty, and Flynn (1999). In general, major clustering algorithms fall into the following classes:

- **Partitioning algorithms:** Given a set of *n* objects, a partitioning algorithm creates k, $k \leq n$, clusters. Such an algorithm usually starts with some initial partitioning and then iteratively improves it by reallocating objects from one cluster to another. Typical examples of partitioning algorithms are the *k*-means algorithm and the *k*-medoids algorithm.
- **Density-based algorithms**: Many partitioning algorithms cluster objects based on the distance between them, thus allowing only for spherical-shaped clusters. Density-based algorithms, on the other hand, can discover clusters of arbitrary shape. The clusters are continuously enlarged until their density exceeds some predefined threshold. DBSCAN and OPTICS are examples of density-based clustering algorithms.
- **Hierarchical algorithms:** These algorithms could be either top-down or bottom-up. A bottom-up (or agglomerative) algorithm starts with each object forming a separate cluster. Then the algorithm successively combines clusters into larger ones until some termination condition is met. A top-down (or divisive) algorithm starts with all objects forming one cluster and successively splits the clusters into smaller clusters. AGNES and DIANA are, respectively, an agglomerative and a divisive algorithm.
- **Grid-based algorithms:** These algorithms divide the object space into a finite number of cells forming a grid structure. The clustering is performed on the grid. Typical examples of grid-based algorithms include CLIQUE and STING.

Genetic Algorithms

Genetic algorithms and evolutionary programming are algorithmic optimization techniques inspired by natural biological evolution. A genetic algorithm starts with a population of individuals and successively applies selection, crossover, and mutation operations to maximize a fitness function. Each individual is usually represented as a string over a finite alphabet, and every string has its own fitness. In the crossover operation, substrings from a pair of individuals are swapped to form a new pair. In the mutation operator, randomly selected bits in an individual's string are changed. From every population, a new population is formed which consists of the fittest individuals in the population and the offspring of these individuals, produced by crossover and mutation operators. The process of generating populations based on previous populations continues until no further improvements are possible or some termination criterion is met. Genetic algorithms are easily parallelizable and can be used for both classification and optimization problems. The major drawback of genetic algorithms is that they may get stuck on local optima and never find the best solution.

WEB MINING

The term "Web mining" usually refers to the overall process of discovering potentially useful and previously unknown information from Web documents and services. Web mining could be viewed as an extension of standard data mining to web data (Kosala & Blockeel, 2000). Compared with standard data mining, Web mining differs in the type of data, the way the data is collected, and the nature of the patterns discovered. Good reviews of Web mining are provided by Mena (1999) and Srivastava, Cooley, Deshpande, and Tan (2000). Web mining could be divided into three areas: web usage mining, web content mining, and web structure mining.

Web Data

Web mining data can be collected at the server-side, client-side, proxy servers, or obtained from corporate databases. Most of the data comes from the server log files. Every time a user requests a Web page, the Web server enters a record of the transaction in a log file. Records are written in a format known as the Common Log File format (CLF), which has been standardized by the World Wide Web consortium (W3C). Different servers may extend this format in their own way. The most useful fields of a CLF record are the Internet Protocol (IP) address of the host computer requesting a page, the Hypertext Transfer Protocol (HTTP) request method, and the time of the transaction. The IP address field helps identify individual users, if they use a static IP address. All records with the same IP address are combined to track the user navigation. The HTTP method indicates what the user is looking for and which pages he or she has requested. The time of the transaction helps detect how much time the user has spent on every page.

Although CLF is supported by the majority of log-file analysis tools, it has several drawbacks, the main one being the fixed number of fields. In many cases, it is desirable to record more information or to omit some information to improve efficiency. As a solution to these problems, the W3C has proposed the Extended Log File format (ELF). ELF provides the following additional functionalities:

- Permits control over the data recorded
- Supports needs of proxies, clients, and servers in a common format
- Allows exchange of demographic data
- Allows summary data to be expressed

One of the most useful extensions to the CLF is the referrer field, which identifies the site visited before coming to the current page. Referrer information answers the important question of where customers are coming from. If a visitor is coming from a search engine site, the referrer information can even tell you which key words have been used to find the current page. Knowing the search key words helps management and business analysts better understand the identity of a Web site and the reasons customers are visiting the site.

Although server log files are rich in information, the data is stored at a very detailed level that is difficult for humans to understand. In addition, the size of log files is usually extremely large, often ranging into gigabytes per day. The best way to summarize and analyze log files is with traffic analysis and monitoring software. Log file analysis tools import the information recorded in server log files into a database from which they can generate reports. Such reports usually summarize data and provide quantitative and qualitative analysis. The reports could be fine-tuned to meet different needs. A Web designer, for example, could receive an analysis of how visitors navigate the site, which are the most and the least visited pages, and how much time visitors spend on the Web site or on a particular page. Advertisers can run reports on the number of clicktroughs for a banner or a product, the keywords that bring the most traffic to the Web site, demographic statistics and users' locations, routes visitors make to the Web site, and so on.

The interest in interpreting Web usage data has spawned a large market for log file analysis tools capable of summarizing, analyzing, and visualizing Web usage information. There is a great number of commercial off-the-shelf log file analysis tools such as WebTrends, NetTracker, and Astra. Most of the tools produce reports on factors such as Web site visitor patterns, referring sites, navigation paths, demographics, and cookie information. Some log file analysis packages, such as Astra, provide visualization tools capable of displaying aggregate counts and spatial views of Web traffic over time. More sophisticated visualization tools can summarize large quantities of data, visually orient data from different perspectives, and proactively expose trends.

It is worthwhile to differentiate between log-file analyzers and sniffers. A network sniffer is a system for interception and analysis of packets transmitted through a network. It is usually installed on a local network computer, and it is able to observe all traffic, including packets not addressed to the chosen computer. A network sniffer is an alternative solution to the problem of tracking visitors' behavior. It is especially attractive in cases in which a Web site is deployed on many Web servers, because the sniffer can collect information from all these servers at once.

Another data source for Web mining are cookies containing state related information, such as a user ID, passwords, shopping cart, purchase history, and customer preferences. Cookies help keep track of several customer's visits and build a customer's profile. Some marketing networks, such as Doubleclick, use cookies to track customers across many Web sites. Other tracking devices, currently producing a lot of controversy, are Web bugs, or clear GIFs (graphical interchange format). A Web bug is a hidden image in a Web page that activates a third-party spying devise without being noticed by Web page visitors. Web bugs are usually used for tracking users' purchasing and browsing habits.

Web Usage Mining

Web usage mining is applying the techniques of data mining to Web data to understand and predict user behavior. To build better predictive models, some additional domain knowledge is generally used, such as concept hierarchies, a Web site topology, navigation templates, and so on. Some Web usage mining applications directly access Web data using some preprocessing techniques. Other applications transform and save Web data as relational databases before using it. The main goals of Web usage mining include improving the effectiveness of Web sites, providing personalized Web content, and eliciting users' needs and preferences. Because of the special importance of Web usage mining to e-commerce, we discuss in separate sections two common types of Web usage mining: clickstream analysis and Web site evaluation.

Clickstream Analysis

Clickstream analysis is a special type of Web usage mining that provides information essential to understanding the effectiveness of Web site design, marketing and merchandising efforts. The concept of clickstream usually refers to a visitor's path through the Web site. It contains the sequence of actions entered as mouse clicks, keystrokes, and server responses as the visitor navigates a Web site. Clickstream data can be obtained from a Web server's log files, commerce server database, or from client-side tracking applications. Analyzing clickstream data is of special importance to online marketing and merchandizing. Web metrics have been developed to help evaluate the effectiveness of online marketing. The most common metrics include impression (a measure of how many times a banner is displayed), clicktrough rate (the ratio of the number of times a banner is shown to the number of times it is clicked on), conversion rate (the percentage of visitors who made purchases), and banner ad return on investment (the amount of profit generated by visitors referred by a banner ad). Clickstream analysis tools usually classify hyperlinks by their merchandising or marketing purposes (for example, a hyperlink could be a cross-sell link referring to a complementary product). This helps in tracking and measuring traffic on hyperlinks. Many Web log analysis tools provide standard reports on clickstream patterns that analyze, summarize, and visualize Web usage patterns. One common drawback of log-file analysis tools is the passive interpretation of clickstream data in the form of noninteractive static diagrams and charts. To overcome these limitations of log-file analysis packages, Lee, Podlaseck, Schonberg, and Hoch (2001) developed an interactive visualization system, providing interactive interpretation of clickstream data. The system analyzes the effectiveness of a Web site by using two visualization techniques: visualization of sessions using parallel coordinates and visualization of product performance using starfield graphs. Starfield graphs are general purpose analysis tools for finding patterns in multidimensional data. The starfield display helps identify the relative significance of different products and find out whether they are over- or underexposed on a Web site.

Web Usage Mining for Web Site Evaluation

Early methods for measuring and studying the quality of Web sites were based on interviewing or questioning representative groups of customers. Such methods tend to incur significant overhead in terms of establishing an experimental environment and selecting a representative group. In addition, these methods are difficult to tailor to individual customers and difficult to perform on a regular basis.

Because the activities of all users are recorded in log files and can be easily associated with e-commerce servers' data, data mining provides a natural way to evaluate the quality and measure the success of commercial Web sites. Within the framework of Web-based marketing, two measures of a Web site's success have been traditionally used: contact and conversion efficiency (Spiliopoulou & Pohle, 2001). The contact efficiency is the percentage of visitors engaged in exploring the Web site, whereas the conversion efficiency is the percentage of visitors that purchased something or achieved the site's goal. In addition, the quality of a Web site could be measured on the basis of factors such as response time, quality of navigation, accessibility of a page, time spent, and so on.

To improve their Web sites, some companies employ clustering methods to discover correlated but not linked pages. The results are used to generate dynamic pages, pointing to all related topics within a Web site. Another method uses conjoint analysis to find a Web site's optimal shape from a user's point of view. The principle of conjoint analysis is that different visitors attach values to all the attributes of a Web site. The total value of a Web site is usually the sum of the values of its components. After finding the optimal value and the optimal Web site configuration for a visitor, individualized layouts can dynamically be generated for different users.

Another common technique is to find the most frequent paths in a Web site and to suggest them to new and inexperienced visitors. Most frequent navigation paths show general tendencies when browsing a Web site, and they are usually extracted using association rule discovery or sequence mining. It is of significant importance to the success of the site that the most frequent paths are optimized with respect to the site's business goal. The IBM SurfAID goes one step further by trying to discover the evolution of navigation patterns for a particular time spot, using association rules and time series analysis.

The Web Utilization Miner (WUM; http://wum.wiwi.hu-berlin.de) is a Web-mining tool especially designed for analyzing and improving Web sites. WUM differentiates between action pages and target pages. Action pages are defined as pages invocation of which indicates that the user is pursuing the site's goal (purchasing a product, providing information, etc.). Target pages are pages on which the site's goal is usually achieved. WUM identifies all pages with low contact efficiency, allowing for improving the structure of a Web site. In addition, it analyses the conversion efficiency for all active sessions to identify pages that cause difficulties for visitors.

Web Content Mining

Web content mining relates to the discovery of useful information from the Web documents, which may consist of text, audio, video, image, and metadata. A subfield of data mining, called multimedia data mining, specializes in mining multiple types of data. Another data mining subfield, termed knowledge discovery in texts, deals with semistructured or unstructured data, such as HTML documents. Most applications in the area of Web content mining represent unstructured documents as vectors of words and their frequencies (Baeza-Yates & Ribeiro-Neto, 1999). Some preprocessing, such as Latent Semantic indexing, is usually applied to reduce the dimensionality of the vector space. The applications in Web content mining range from text classification and categorization, event detection and tracking, and text clustering, to finding sequential patterns and associations in text documents. Topic detection and tracking (TDT) is an interesting new line of research in Web content mining (Allan, Carbonell, Doddington, Yamron, & Yang, 1998). Topic tracking is the process of monitoring a stream of news stories to find which of them are tracking the same event. Topic detection involves detecting the occurrence of a new event such as stock price or interest rate change, in a stream of news stories from multiple sources.

Web Structure Mining

Web structure mining tries to discover the network structure and the topology of the Web hyperlinks. It is especially useful for categorizing Web pages and for finding hidden relationships between Web sites. PageRank was one of the first algorithms measuring the relative importance of web pages based on the graph structure of the Web. The algorithm was developed by Sergey Brin and Lawrence Page and is used by Google to rank Web pages based on their relevancy. The rank of a Web page is defined as a weighted sum of the ranks of all pages pointing to the page. In another interpretation, the rank of a page is the probability that a random surfer visits the page.

Another page ranking algorithm, called hubs/authorities model, was proposed by Kleinberg. The hubs/authorities was inspired by social network analysis and discovers specific types of pages based on their incoming and outgoing links. In the hubs/authorities model, pages with many links pointing to them are called authorities in the sense that they are considered to deliver relevant content. Pages with many outgoing links are called hubs in the sense that they are considered to point to similar content. The main idea behind the hubs/authorities model is that good hubs point to good authority pages, and good authority pages have been pointed to by good hubs.

DATA MINING AND PERSONALIZATION

Businesses use personalization techniques to increase sales, perform target advertising, improve customer service, make Web sites easier to use, increase customer base, develop customer loyalty, and so on. Database marketing, for example, makes extensive use of personalization techniques to market a particular product to a specific customer or group of customers. In this section, we discuss two major personalization areas: customer profiling and recommender systems.

Customer Profiling

One of the key issues in developing e-commerce applications is the construction of profiles for individual customers, providing information about who these customers are and how they behave. Electronic profiling is based on the customers' online behavior and transactional histories, which can be captured by registration forms, log files, cookies, and collaborative software. Most of the work on individual profiling is proprietary and has been done in the industry.

A profile is a collection of data describing an individual user or a group of users. There are two main types of customer profiles: factual and rule-based (Adomavicius & Tuzhilin, 2001). A factual profile usually contains demographic attributes (e.g., address, education, gender, income level), and some additional transactional data (e.g., maximal purchase amounts for a customer, favorite products). Factual profiles have been used by Engage Technologies (http://www.engage.com). Whereas factual profiles describe who the customer is, rule-based profiles describe what the customer does. Such profiles usually contain association rules describing the customers' behavior. The rules can be defined by a human expert or extracted from transactional data using data mining methods. Rule-based profiles have been used by Broad Vision (http://www.broadvision.com) and Art Technology Group (http://www.atg.com).

The rule-based profile-building process usually consists of two main steps: rule discovery and rule validation. Various data mining algorithms can be used for rule discovery, such as Apriori, FP-Growth, and CART. A special type of association rules, profile association rules, has been proposed by Agrawal, Sun, and Yu (1998). A profile association rule is one in which the left-hand side consists of customer profile information (e.g., age, salary, education). The right-hand side includes customer behavior information (e.g., buying beer, using coupons). Agrawal et al. (1998) proposed a multidimensional indexing structure and an algorithm for mining profile association rules. Profile association rules are especially useful for segmenting users based on their transactional characteristics and for deriving customers' behavioral attributes from their transactional attributes.

One of the problems with many rule discovery methods is the large number of generated rules, many of which, although statistically acceptable, are spurious, irrelevant, or trivial. Postanalysis is usually used to filter out irrelevant and spurious rules. Several data mining systems perform rule validation by letting a domain expert inspect the rules on one-by-one basis and reject unacceptable rules. Such an approach, however, is not scalable to large numbers of rules and customer profiles. Another solution was proposed by Adomavicius and Tuzhilin (2001). The idea behind the solution is that many users share common or similar rules that can be validated together. Rules are collected into one set, and several rule validation operators are applied iteratively to the set. The collective rule validation allows a human expert to reject or accept a large number of rules at once, thereby reducing validation effort.

Another interesting application of user profiling is personalized Web search and information retrieval. A personalized Web browser is capable of learning user's access behavior. A typical example is Letizia (http://www.media.mit.edu), which monitors the user's browsing behavior, develops a user profile, and searches for potentially interesting pages. The user profile is developed in nonintrusive ways while the user is browsing the Web. Another approach is to seek the user's help and ask the user to rank pages. Based on the rankings and the content of the pages, a browser can learn a user profile and help predict which pages are of interest to the user.

Recommender Systems

Recommender systems (Schafer, Konstan, & Riedl, 2001) have been used in B-to-C e-commerce sites to make product recommendation and to provide customers with more information to help them decide which product to buy. Recommender systems provide a solution to the problem of how to choose a product in the presence of overwhelming amount of information. Many e-commerce sites offer millions of products, hence choosing a particular product requires processing of large amounts of information, thereby making consumer choice difficult and tedious.

Recommender systems contribute to the success of e-commerce sites in three major ways (Schafer et al., 2001). First, they help improve cross-sell. Cross-sell is usually improved by recommending additional products for the customer to buy. For example, by looking at the products in the customer's shopping cart during the checkout process, a system could recommend additional complementary products. Second, recommender systems could help convert occasional visitors into buyers. By providing a recommendation, a retailer could deliver customized information, increase the amount of time spent on a Web site, and, finally, increase the customer's willingness to buy. Third, recommender systems help build loyalty and improve customer retention. Personalized recommendations create a relationship between a Web site and a customer. The site has to invest additional resources into learning customers' preferences and needs; and customers have to spend time teaching a Web site what their preferences are. Switching to a competitor's Web site becomes time-consuming and inefficient for customers who have to start the process of building personalized profiles from the beginning. In addition, customers tend to return to Web sites that best match their needs.

Recommender systems use different methods for suggesting products (Schafer et al., 2001). One of the most common is item-to-item correlation. This method relates one product to another using purchase history, attribute correlation, and so on. For example, CDNOW suggests a group of artists with similar styles to the artist that the customer likes. Another recommendation method is user-to-user correlation. This method recommends products to a customer based on a correlation between that customer and other customers visiting the same Web site. A typical example is "Customers who bought." When a customer is buying or browsing a selected product, this method returns a list of products purchased by customers who purchased the selected product. Other recommendation methods include statistical summaries (Top-10 list), expert reviews, and customers' comments, among other information.

Collaborative filtering (CF) was one of the earliest recommender technologies. CF systems make a recommendation to a target customer by finding a set of customers, called a neighborhood, that have similar taste with the target customer. The neighborhood includes customers who have a history of agreeing with the target user. Products that the neighbors like are then recommended to the target customer. In other words, CF is based on the idea that people who agreed on their purchasing decisions in the past are likely to agree in the future. The process of CF consists of the following three steps: representation of products and their rankings, neighborhood formation, and recommendation generation.

During the representation stage, a customer-product matrix is created, consisting of ratings given by all customers to all products. The customer-product matrix is usually extremely large and sparse. It is large because most online stores offer large product sets ranging into millions of products. The sparseness results from the fact that each customer has usually purchased or evaluated only a small subset of all products. To reduce the dimensionality of the customer-product matrix, different dimensionality reducing methods can be applied, such as latent semantic indexing and term clustering.

The neighborhood formation stage is based on computing the similarities between customers and grouping like-minded customers in one neighborhood. The similarity between customers is usually measured using either the correlation or the cosine measure. After computing the proximity between customers, a neighborhood is formed using clustering algorithms.

The final step of CF is to generate the top-N recommendations from the neighborhood of customers. Recommendations could be generated using the most frequent item technique, which looks into a neighborhood and sorts all products according to their frequency. The *N* most frequently purchased products are returned as a recommendation. Another recommendation technique uses the discovery of association rules. It finds all rules supported by a customer, that is, the customer has purchased the products from the left-hand side of the rule. Then the products from the right-hand side of the rule are returned as a recommendation.

Another recommendation technique is content-based recommendation (Balabanovic & Shoham, 1997). Although collaborative filtering identifies customers whose tastes are similar to those of the target customer, content-based recommendation identifies items similar to those the target customer has liked or purchased in the past. Content-based recommendation has its roots in information retrieval (Baeza-Yates & Ribeiro-Neto, 1999). For example, a text document is recommended based on a comparison between the content of the document and a user profile. The comparison is usually performed using vectors of words and their relative weights. In some cases, the user is asked for feedback after the document has been shown to him. If the user likes the recommendation, the weights of the word extracted from the document are increased. This process is called relevance feedback.

However, content-based recommender systems have several shortcomings. First, content-based recommendation systems cannot perform in domains where there is no content associated with items, or where the content is difficult to analyze. Second, only a shallow analysis of very restricted types of content is usually performed. To overcome these problems, a new hybrid recommendation technique has been proposed (Melville, Mooney, & Nagarajan, 2002). It is called content-boosted collabo rative filtering. The technique uses a content-based predictor to enhance existing user data and then provides a recommendation using collaborative filtering.

DATA MINING IN WEB MARKETING

Data mining applications help marketers discover new behavioral and attitudinal customers' characteristics. The four main Internet marketing activities include discovery and analysis of customer attraction, sequential pattern discovery for customer retention, rule induction, clustering for cross-selling, and preempting departure (Buchner, Anand, Mulvenna, & Hughes, 1998).

Customer Attraction

The three major goals of customer attraction analysis are as follows: selection of new prospective customers, acquisition of selected customers, and dropping out existing nonprospective customers. Web site visitors are usually grouped into clusters: no customers (visitors who never made a purchase), irregular customers, and regular customers. Association rules could be used for describing the behavior of each group of customers. To increase the conversion rate of irregular customers, or to retain regular customers, dynamically generated Web content could be displayed to each group.

Customer Retention and Sequential Patterns

In e-commerce, where competitors are only one click away, retaining customers is an extremely challenging task. Sequential pattern discovery helps marketers find intertransaction and navigation patterns across time. In intertransaction patterns, the presence of a set of items is followed by another item within a particular time interval. Navigation patterns are similar to intertransaction patterns with the only difference that they capture sequence of frequently visited pages. The techniques used for sequential pattern discovery are usually based on association rule discovery and cluster analysis. Data is usually collected from transactional databases and from the Web server's log files. The discovery of visitors' sequential patterns allows marketers to predict users' visits, and purchasing and browsing patterns, and helps them target advertising, such as banner selection and e-mail marketing.

Cross-Selling

The goal of cross-selling is to extend horizontally or vertically the purchasing activities of existing customers. A common data mining method for cross-selling is attribute-oriented induction based on a hierarchy of Web pages. Attribute-oriented induction is a learning-by-example technique that extracts generalized data from databases. When a user visits a page, the page is replaced by its corresponding more general page from the hierarchy, thereby providing the visitor with more choices.

Customer Departure

The goal of customer departure analysis is to predict and prevent customer exit by taking appropriate action, such as targeted promotion or content targeting. Customers depart from a Web site when they either move to a competitor or stop purchasing a certain product or service. Because a customer departure is not directly observable, a use-defined threshold interval is chosen, in which no activities have been recorded. Log files from certain period to the last activity are analyzed to find the characteristics of the customers who left the Web site.

DATA MINING IN MARKET BASKET ANALYSIS

The goal of market basket analysis is to determine which products are purchased together by a customer, that is, which products a customer puts into his or her shopping cart or "market basket." Market basket analysis provides several advantages to businesses. First, it helps retailers better design store layout or online catalogs by combining and displaying together all products that are likely to be purchased together. Second, market basket analysis helps direct marketers target particular customers based on their purchasing histories. Third, knowing which products sell together can help inventory management.

A special type of market basket analysis is differential analysis. In differential analysis, data mining analysts compare results, for example, between different stores, between customers in different segments, or different seasons of the year. Differential analysis helps analysts filter out spurious patterns and discover patterns that apply to specific subsets of data. Rule induction is usually used for market basket analysis. Most commercially available data mining systems, such as Nuggets, IBM Intelligent Miner, and Clementine, provide tools for mining association rules.

Market basket analysis can be applied outside retail industry. Some other applications include the following:

- **Analysis of credit card purchases:** Items purchased using a credit card can be used to predict future purchases.
- **Analysis of telephone calling patterns:** By identifying which services tend to be purchased together by telecommunication customers, companies can bundle these services together and increase revenue.
- **Identification of fraudulent medical insurance claims:** Unusual combinations of medical insurance claims can signal fraudulent behavior.
- **Analysis of banking services:** By identifying specific services (car loans, certificates of deposit, etc.) that are likely to be purchased together, banks can offer customers other services that are likely to be purchased together with the services customers already purchased.

CONCLUSION

E-commerce companies can greatly benefit by deploying data mining technologies. The fast-growing e-commerce data exceeds human abilities for data collection, analysis, and decision making. Web mining can help businesses reveal important data patterns, build customer profiles, improve the efficiency of Web sites, and perform controlled experiments, among other things. Data mining helps address the fundamental issue of personalization by providing solutions for understanding customers' behavior, and generating personalized content. To meet the constantly increasing information needs of the highly competitive e-commerce markets, new techniques for collection and analysis of Web data have been developed, such as log-file analysis and recommender systems.

Because of the large variety of data mining applications in e-commerce, this overview is by no means exhaustive. There are many other data mining applications in e-commerce including applications in financial modeling and fraud detection. Financial modeling uses predictive models, such as artificial neural networks and regression, for portfolio optimization, investment selection, and trading models. Carlberg & Associates, for example, developed a neural network model for predicting the Standard & Poor's 500 Index, using interest rates, dividends, earnings, and oil prices. LBS Capital Management, an *Inc.* 500 fund-management company, uses expert systems, neural networks, and genetic algorithms for asset management and risk analysis. The FALCON fraud assessment system from NHC uses neural network to detect suspicious credit card transactions. Many retail banks use it for fraud protection.

Although promising, data mining in e-commerce is still in its infancy. The Web presents new problems and challenges to traditional data mining. The main obstacles to successful mining of e-commerce data are as follows:

- **The unstructured web data:** Web data is usually stored in different formats on different media.
- **Low-quality data:** Web server logs cannot accurately identify sessions or users. It is usually difficult to detect critical events using server logs.
- **Scalability:** Most Web data mining algorithms cannot process the amount of data generated at Web sites.
- **Data integration:** Integration of Web usage data with Web content and Web structure data. Integration of data from various sources in different formats.

There are several directions for future research and development. An interesting direction for e-commerce data mining is discovering unexpected information, that is, information that cannot be anticipated by the user. Most of the data mining applications essentially look for information matching some previously established expectations in the form of predefined models. Recent work by Liu, Ma, and Yu (2001) proposes data mining methods that help find unexpected information from competitors' Web sites. Unexpected information may include unknown services, products, or information that competitors have. With this information, a company can design countermeasures to improve its competitiveness.

Topic detection and tracking (Allan et al., 1998) is another promising direction for e-commerce data mining. TDT is a Defense Advance Research Projects Agency (DARPA) sponsored initiative that finds and tracks new events in a stream of broadcast news stories. TDT could help businesses detect the appearance of new topics, such as the release of a competitive product, price dynamics, customers' feedback, or financial news, and track their evolution.

The main promise of data mining technology is its ability to cope with and exploit the ever-growing amount of online data. As competitive pressures increase, data mining is going to play an ever-increasing role for the success of e-business.

GLOSSARY

Artificial neural network A nonlinear technique based on a model of a human neuron; a neural net is used to predict outputs from a set of inputs.

Association rule An if–then rule describing how often certain items have occurred together.

Clustering The process of grouping the data into clusters based on the principle of maximizing intraclass similarity and minimizing interclass similarity.

Collaborative filtering Recommendation technology that makes a suggestion to a target customer by finding a set of customers that have similar taste.

Content-based recommendation Recommendation technology that makes a suggestion to a target customer by finding items similar to those the target customer has liked or purchased in the past.

Customer profile A collection of data that describes an individual user or a group of users.

Data mining The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data (Fayyad et al., 1996).

Decision tree A tree representing a collection of tests that lead to a class or a value.

Genetic algorithm An optimization algorithm based on natural evolution concepts such as genetic combination, mutation, and natural selection.

Log file A file generated by Web servers that keeps a record for every transaction.

Personalization The use of technology and customer information to tailor e-commerce interactions between a business and each individual customer.

CROSS REFERENCES

See *Data Warehousing and Data Marts; Databases on the Web; Intelligent Agents; Machine Learning and Data Mining on the Web; On-Line Analytical Processing (OLAP); Personalization and Customization Technologies.*

REFERENCES

Adomavicius, G., & Tuzhilin, A. (2001). Expert-driven validation of rule-based user models in personalization applications. *Data Mining and Knowledge Discovery, 5,* 33–58.

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. *Proceedings of the International Conference on Very Large Data Bases (VLDB'94)* (pp. 487–499). San Francisco: Morgan Kaufmann.

Agrawal, C., Sun, Z., & Yu, P. (1998). Online generation of profile association rules. *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining* (pp. 129–133). Menlo Park, CA: AAAI Press.

Allan, J., Carbonell, J., Doddington, G., Yamron, J., & Yang, Y. (1998). Topic detection and tracking pilot study: Final report. *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop* (pp. 194–218). Gaithersburg, MD: NIST.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval*. Boston: Addison Wesley.

Balabanovic, M., & Shoham, Y. (1997). Combining content-based and collaborative recommendation. *Communications of the ACM, 40*(3), pp. 66–72.

Buchner, A., Anand, S., Mulvenna, M., & Hughes, J. (1998). Discovering Internet marketing intelligence through online analytical web usage mining. *ACM SIGMOD Record, 27*(4), 54–61.

Han, J., Pei, J., & Yin, Y. (2000). Mining frequent patterns without candidate generation. *Proceedings of ACM-SIGMOD international conference on management of data.* ACM, pp. 1–12.

Fayyad, U., Piatesky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). *Advances in knowledge discovery and data mining.* Cambridge, MA: MIT Press.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys, 31,* 264–323.

Kohavi, R., & Provost, F. (2001). Applications of data mining to electronic commerce. *Data mining and knowledge discovery, 5,* 1–7.

Kosla, R., & Blockeel, H. (2000). Web mining research: A survey. *SIGKDD Explorations, 2,* 1–15.

Lee, J., Podlaseck, M., Schonberg, E., & Hoch, R. (2001). Visualization and analysis of clickstream data of online stores for understanding web merchandising. *Data mining and knowledge discovery, 5,* 59–84.

Liu, B., Ma, Y., & Yu, P. (2001). Discovering unexpected information from your competitors' Web sites. *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 144–153). New York: ACM.

Melville, P., Mooney, R., & Nagarajan, R. (2002). Content-boosted collaborative filtering for improved recommendations. *Proceedings of the 18th National*

Conference on Artificial Intelligence (pp. 187–192). Menlo Park, CA: AAAI Press.

Mena, J. (1999). *Data mining your website.* Boston: Digital Press.

Schafer, J., Konstan J., & Riedl, J. (2001). E-commerce recommendation applications. *Data Mining and Knowledge Discovery, 5,* 115–153.

Spiliopoulou, M., & Pohle, C. (2001). Data mining for measuring and improving the success of Web sites. *Data Mining and Knowledge Discovery, 5,* 85–114.

Srivastava, J., Cooley, R., Deshpande, M., & Tan, P. (2000). Web Usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explorations, 1*(2), 1–12.

Data Warehousing and Data Marts

Chuck Kelley, *Excellence In Data, Inc.*

INTRODUCTION

Data warehousing and data marts bring a new way for organizations to learn about their internal and external customers, vendors, and employees. The problem is that while tremendous amounts of data have been captured, there has not been an environment in which the user community could share the knowledge gained from and with their customers, vendors, or employees. With the advent of the Internet, ways must be devised to capture data, transform that data into information, and provide access to that information.

The Internet provides a strong environment where the organization can share information within the user community, whether they are employees, vendors, or customers. The sharing of information can provide value to their customers, thereby becoming a competitive weapon. This competitive weapon will help organizations to grow.

Information technology (IT) organizations have embarked on data warehousing and data marts as a method of accomplishing the transforming of data into information and have been providing tools to allow the knowledge workers to "mine" the data to find knowledge. Once this knowledge has been "mined," there must be a process in place to make sure that this new knowledge is true. Once it is determined to be true, then the new knowledge must be fed to the operational systems and the user community. The highly generic formula (Kelley, 2001)

$$\text{Value of Data Warehouse} = \text{Quality of Knowledge} \times \text{Organizational Reach}$$

is a strong requirement for data warehousing on the Internet. In this formula, the more people that learn about high-quality knowledge, the more value the data warehouse adds to the organization. The Internet provides the wide organization reach to that knowledge. The value of the data warehouse will vary, based on each organizations' ability to change, the subject areas defined, and the quality of the data, to name a few.

How do organizations get to this environment? By building a data warehouse architecture that focuses on

Capturing data;
Cleaning the data (thereby providing high-quality data);
Transforming that data into information; and
Delivering the data to the user community.

THE DATA WAREHOUSING PRIMER

Where Did the Data Warehouse Come from?

Data have been captured by transaction systems since the beginning of the computing age. Until recently, there has been no easy way to get the captured data to the user community. Giving access directly to the transaction systems has proven to be bad for performance and the results may not be high-quality or integrated data. Examples of why this is true are below:

More data remain in the transaction system for no apparent reason other than that is where they are stored today (bad performance).

It is hard to tune a system to do both transactions (four reads and five writes) and analysis (read a million rows and aggregate) (bad performance).

Transaction systems are about capturing data, not capturing the highest level of quality data (not high-quality data).

Transaction systems are about doing one job, not worrying about whether they capture the exact same data that another transaction system captures just so that they can be tied together (not integrated).

The data warehouse was developed to solve the needs of the user community concerning these issues.

Data warehousing provides an integrated, historical collection of data. Historical information allows the user community to do trend analysis over a long period of time. There is no longer a need to keep all the history in the transaction systems, unless it is required by the application. This makes them leaner and meaner, so that they can perform like they should.

By taking all the data from the transaction systems and putting it in an integrated collection of data designed for the user community, the data warehouse provides the much needed holistic view of the customer and their relationship(s) with the organization.

The Data Warehouse Environment

Data warehousing is an environment that allows user community to access data that have been structured for query. Querying is asking questions and getting results that will allow the user to "perceive" information that will be more valuable to the user community. Sometimes this can be a multi-step process. Think about how decisions are made. An "Ah, Ha!" is the ultimate goal. "Ah, Ha!" can be defined as finding something not known before that may be of value to the user community.

This "Ah, Ha!" might take a significant number of steps. The first query might start with a tremendous number of rows. An anomaly might be perceived in the data. Looking at that subset of data might lead the user down a path that may or may not be fruitful. (However, one could say that it will always be fruitful because information was learned that was not known previously!) If not, then back up to the previous data set and try another path through the data. Eventually, some knowledge will be gained from this process of thought. That knowledge might be as simplistic as "the data shows no trend" toward your hypothesis or "there is a relationship between abnormally large purchases (using credit cards) of clothes and the likelihood of that person declaring bankruptcy."

Data warehousing, if built properly, provides the environment to peruse the data to find new knowledge and generate value.

The data warehousing environment has been through a lot of changes since the early 1990s. The largest areas of change are in (1) the implementation of the data warehousing environment and (2) accessing the data in the data warehouse via the Internet. There are two major concepts strongly associated with the implementation of a data warehousing environment—(1) the data warehouse and (2) data marts.

Data Warehouse

The data warehouse is defined as a (1) subject-oriented, (2) integrated, (3) time variant, (4) nonvolatile collection of data in support of management's decisions (Inmon, 1992).

Subject oriented—Subject oriented means that each area of the data warehouse is built to solve a business problem. That problem can be (but is not limited to) (1) understanding your customer's buying habits, (2) processes that affect the quality of your product, or (3) efficiency of your staff. Subject areas have no application flavor at all. In fact, they will typically take data from multiple applications to create a single subject area. For example, a bank may create a customer data warehouse that contains transactions from the checking accounts, saving accounts, loan department, and investments. While each has its own application, the subject area within the data warehouse would look nothing like those applications.

Integrated—Integrated means that there is a single uniform representation of data stored in the data warehouse. If some of the manufacturing plants keep track of copper wire in centimeters, others in millimeters, and others in inches, there needs to be a uniform way of storing this in the data warehouse. That way, analysis can be done without having to remember that unit of measure each plant uses, or worse yet, adding quantities together without realizing the units of measure are different. Integration also allows the organization to implement their standardized definitions.

Time variant—All data as they enter the data warehouse are accurate at some point in time. Associating all pieces of data with a moment in time allows the user community to understand trends, correlations, etc. over time.

Nonvolatile—Nonvolatile is a synonym for read-only. Having data change in the data warehouse means that we can restate history—never (OK, never say never!) a good thing to do. If data are valid as of one moment, they need to be associated with that moment of time in the data warehouse.

While the data warehouse is always shown as a large monolithic database, it can be a distributed database (Inmon & Kelley, 1993). Distributed databases add a level of complexity in the creation of the data warehouse. This complexity includes the following:

The coordination of extracts of the source systems;
The use of different algorithms for calculations; and
The synchronization of loads into the data warehouse.

During the early days, there were introductions to the data warehouse defined as wholesale/retail models of data warehouse. The wholesale data warehouse was in fact, conceptually, a monolithic data warehouse. The retail data warehouse was a subset of the data warehouse, fed from the data warehouse, but containing only the data required for that business area. Another way the data warehouse was described is that the enterprise data warehouse feeds the departmental data warehouse that feeds the personal data warehouse. What happened over time is the renaming of these fed systems (retail, departmental, and personal data warehouses) as data marts.

The task of building the enterprise data warehouse has become unyielding. The pressure to get something out to the user community became quite overbearing. Therefore, there is a push to build data marts. The term data mart can be described as avoiding the impossibility of tackling the enterprise data warehouse planning job all at once (Kimball et al., 1998).

Data Marts

Data marts came into place originally as a way to describe a subset of data produced from the data warehouse. The data mart concept has grown to be a lot more than that.

Now, the data mart is built without the data warehouse. The question that must be asked is, "How is the plethora of data marts being controlled?" Most data marts are being built without a framework of the concept of the data warehouse in mind. A mechanism to tie the data marts together is needed—conformed dimensions (Kimball et al., 1998) (dimensions will be discussed in the section Creating the Data Model). A conformed dimension is data that have the exact same meaning as that used in multiple data marts. Therefore a customer dimension in one data mart would closely resemble a customer dimension in a second data mart. Without the concept of conformed dimensions, stovepipe data marts will most likely be created. Stovepipe data marts will not allow the user community to do cross-data mart analysis, and hence, cross-business analysis.

With the exception of the enterprise data warehouse-only environment (see the next section), a data mart is where the user community accesses the data. Data marts are created to do specific studies using on-line analytical processing (OLAP) techniques including ROLAP (relational OLAP), HOLAP (hybrid OLAP), DOLAP (desktop OLAP), MOLAP (multidimensional OLAP), and data mining. OLAP and data mining are discussed elsewhere in the Internet Encyclopedia.

Architectures of the Data Warehouse Environments

There are three primary architectures of the data warehousing environments. They are as follows:

Enterprise data warehouse-only (Figure 1);

Enterprise data warehouse feeding data marts (Figure 2); and

Collection of data marts (Figure 3).

Enterprise Data Warehouse-Only

The enterprise data warehouse-only environment is created with a single data warehouse database. The user community will query directly to the data warehouse for the business needs. The pros of this environment are that it

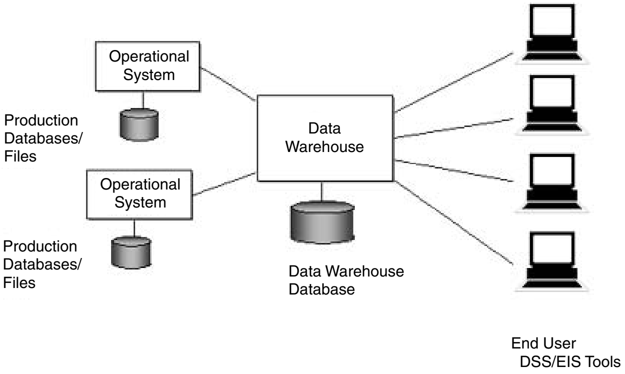


Figure 1: Enterprise data warehouse.

Is the easiest to implement and understand;

Provides for a single point of authority of the data;

Is the easiest to synchronize; and

Provides for a single query point of data for the user community.

The cons for this environment are that

It may need lots of network bandwidth since the user community can be everywhere;

Your database vendor needs to support large tables with lots of parallelism; and

The environment will eventually require large time windows for the extraction, transformation, and load processes (discussed in the section The Extraction, Transformation, and Load (ETL) Processes).

Enterprise Data Warehouse Feeding Data Marts

This environment generally allows access to the data warehousing environment via data marts only. The enterprise data warehouse keeps the data accurate and feeds the data to the data marts. The pros of this environment are that

It gets data closer to the user;

It has a faster response with smaller data sets;

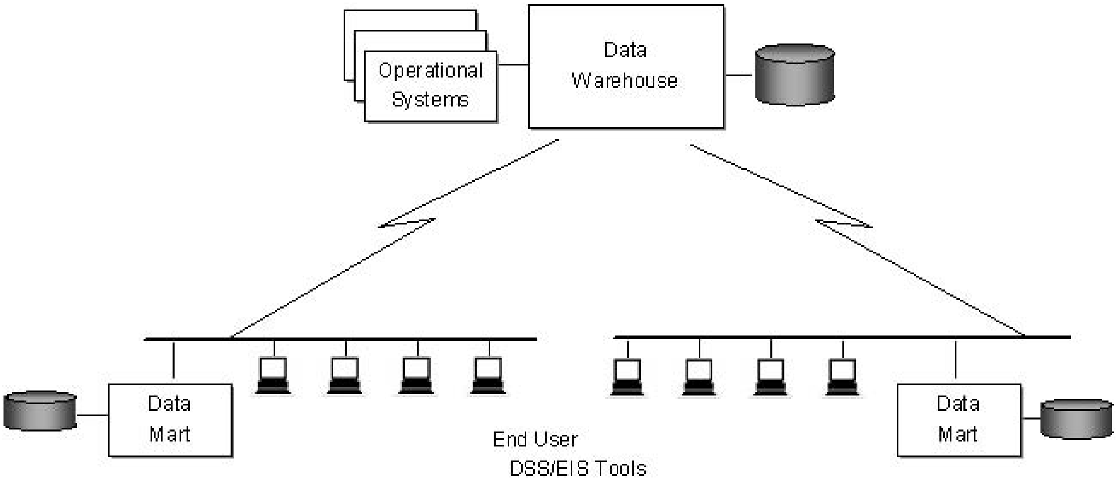


Figure 2: Enterprise data warehouse feeding the data marts.

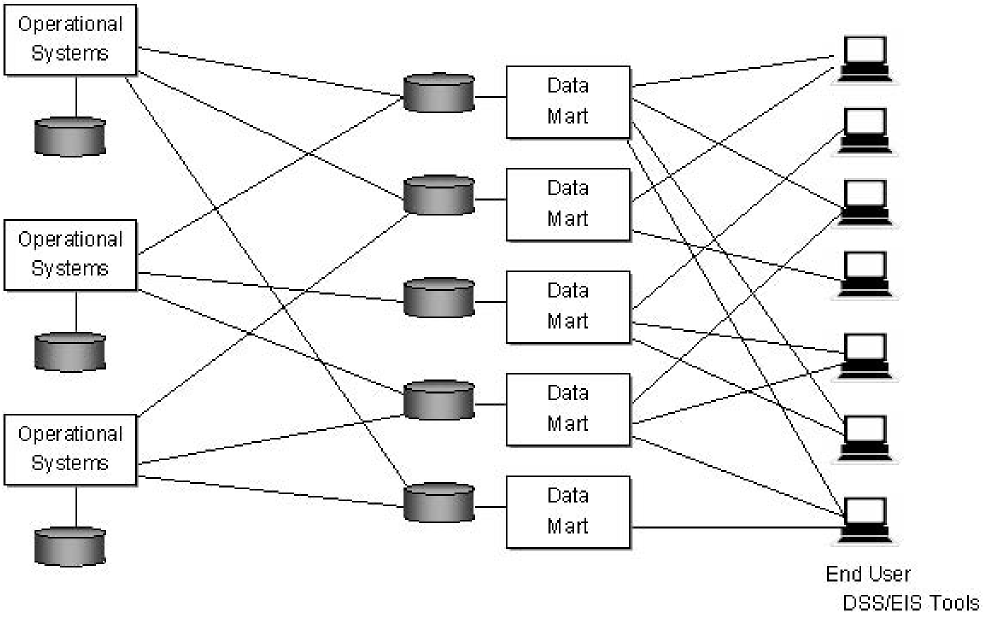


Figure 3: Data warehouse as a collection of data marts.

It can be more finely tuned to usage patterns; and

Data marts are fed from a single authority of data.

The cons of this environment are that

It requires more CPU and disk resources;

Depending on how many data marts, more network bandwidth may be required to move the data; and

Depending on how many data marts, a longer time window for building the data marts will occur because of pushing more data out to data marts.

Collection of Data Marts

A collection of data marts is probably the most complex of the environments to create properly and the one most pushed onto organizations by accountants (money people), consulting organizations, and vendors. Most consulting organizations and vendors offer a "build a data mart in 90 days"-like promotion. If everyone in your organization built data marts without coordinating the efforts, stovepipe data marts will most likely be the outcome.

The pros for this environment are

Smaller, more manageable data sets;

It allows for quick implementation; and

It is generally within someone's budget to "purchase."

The cons for this environment are

Without conformed dimensions, each data mart will have its own reality and will not provide an enterprise-wide view; and

This environment is the most complex to implement due to synchronization and integration issues.

BUSINESS DRIVERS OF THE DATA WAREHOUSE

The business problem that data warehousing first tried to solve was a simple major economic principle—competition. It is competition that is at the heart of business decisions. As the global economy continues to emerge, the information economy will come to an end. Product differentiation may no longer form the basis of competitive advantage. Capturing data, transforming data into information, discovering knowledge from that information, and delivering that knowledge across the organization will become important. Why? Because the only sustainable advantage will be creating *knowledge* and using that knowledge to create *value*. The Internet will be critical in the delivering that knowledge and value to the user community.

Other business drivers of the data warehouse can be noncompetitive items, although they could be interpreted as competitive. These items include quality assurance, medical research, fraud detection, and national security. The largest benefit of the data warehouse environment is that it can provide information and knowledge to help the organization create value for its user community.

Information is the link between an organization's business strategy and how to implement it. A data warehouse allows the organization to measure the success of the strategy. The business value drivers are

The user community can manage their business by accessing the critical information when and using whatever method they desire. That leads to the ability to generate more value for the user community.

The user community can create reports themselves and will no longer be required to wait long lead times for IT to produce reports for them.

The organization can move from tactical-only environment to a "culture of analysis" environment. No longer will the user community spend all of their time doing the tactical tasks, but will start to perform more strategic tasks.

Through the Internet, the flow of data from one part of the organization to other parts will increase, thereby increasing knowledge and value throughout the organization.

With faster access and better information, organizations can make faster adjustments to the marketplace.

STEPS TO CREATING THE DATA WAREHOUSE ENVIRONMENT

Building the data warehouse environment is an easy to talk about but not trivial process. Care must be given to build it in an iterative fashion. Just because the first step is completed doesn't mean it is finished. It is the starting point for the next iteration. As Heraclitus the Obscure (who was born at Ephesus about 530 BC) said, "There is nothing permanent except change." The user community generally does not know what they want until they see something. Once they see something, it will help them define what they really want. Why? Because they are caught up in dealing with the problems of the tactical data (i.e., they are too busy trying to get the data they need to solve their day-to-day activities) that they cannot think about the strategic needs. Tactical issues must be solved before the user community will think strategically.

There are five steps in creating the data warehouse environment. They are

Create the data model;

Define the metadata world;

Build and execute the extraction, transformation and load processes;

Define the hardware architecture; and

Create the user community's world.

Creating the Data Model

Data models are the schematic to the data warehouse environment. It is a representation of how data relates to other data. There are three major data models/schemas used within the data warehouse environment. They are

Star schema;

Snowflake schema; and

Relational (normalized) schema.

Star Schemas

The star schema was introduced through a process called dimensional modeling. Star schemas have two major classes of data—(1) facts and (2) dimensions. Facts are the metrics or measures of the organization that reflect how the organization is doing. Examples of facts include

Quantity sold,

Total revenue,

Revenue per available room (RevPAR),

Number of calls,

Length of calls,

Discounts,

Number of pay-per-views watched,

Total cost, and

Temperature, volume, and pressure.

Dimensions are how the user community wants to look at the metrics "by." Examples of dimensions include

RevPAR "by" day (fact is RevPAR and dimension is time (day));

Total revenue "by" product, "by" day, "by" customer (fact is revenue and dimensions are product, time (day), and customer);

Profit "by" store "by" product "by" month (fact is profit and dimensions are store, product, and time (month)); and

Top 10 customers "by" quarter (fact is unknown at this point, but most likely total revenue or quantity sold and dimensions are customer and time (quarter)).

Facts are created on the basis of granularity of data. Granularity is the level of depth that a fact is stored. Facts are at the intersection of the dimensions. Figure 4 shows a sample star schema.

In this example the fact can be said to be "On this day (time dimension) this product (product dimension) was sold at this store (store dimension) to this customer (customer dimension) for these metrics (fact)." It is critical that the granularity is always kept intact. Having different granularities cause major data anomalies and produce metrics that cannot be added.

Snowflake Schemas

The snowflake schema is similar to the star schema except the dimensions are broken out (normalized) into separate tables. Figure 5 shows an example of a snowflake schema for a different product dimension (not the same product dimension shown in the star schema).

There are two major reasons data modelers want to snowflake—(1) save disk space and (2) less chance of update anomalies (Kimball, 1996) (the major point in normalization).

Compared to the whole data warehouse or data mart, the fact table is the largest and will make up (generally) 80+% of the total size of the data warehouse or data mart. If there is a 50% savings of the remaining 20% for the dimensions, it still will be negligible. The second reason for the snowflake schemas is the less chance of update anomalies. Since the data warehouse or data marts are nonvolatile (read-only), then that should not be a concern.

There are good reasons for the snowflake schema. If the tool chosen by the user community to access the data mart works better with a snowflake schema, then build that data mart with a snowflake schema.

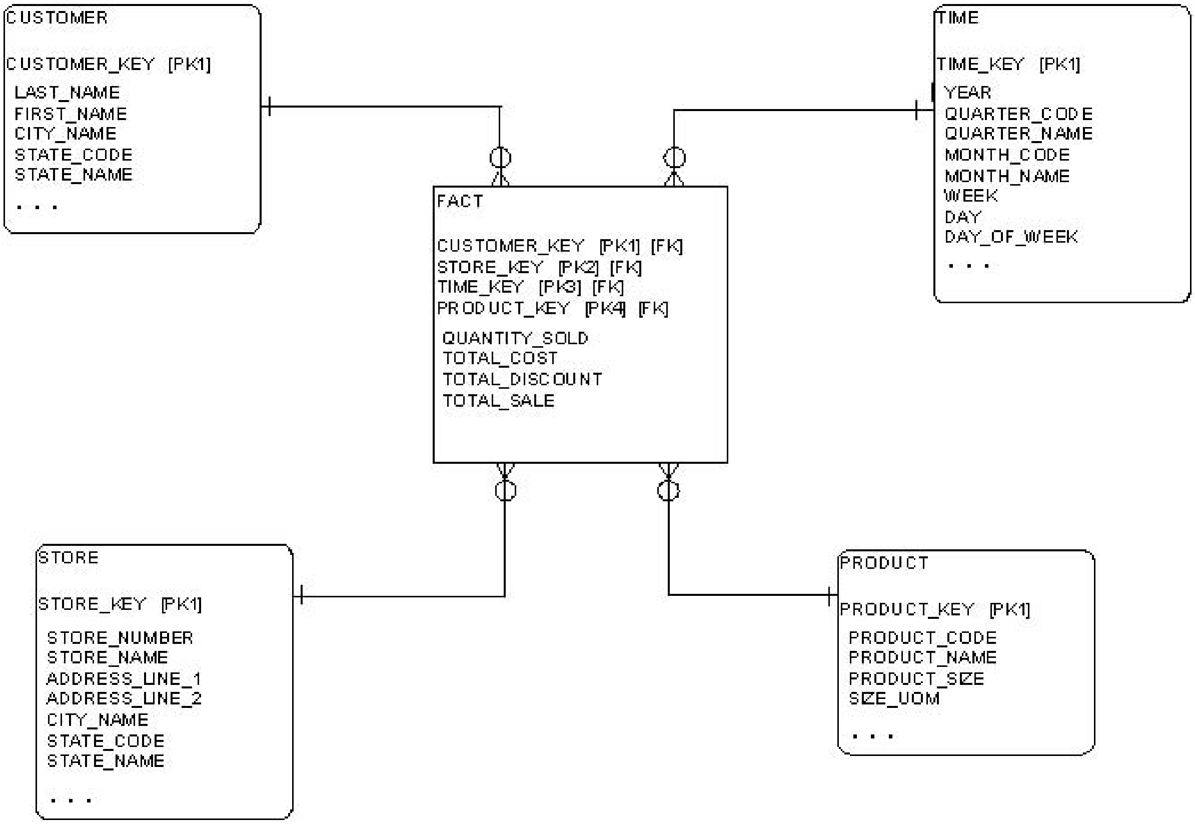


Figure 4: Star schema.

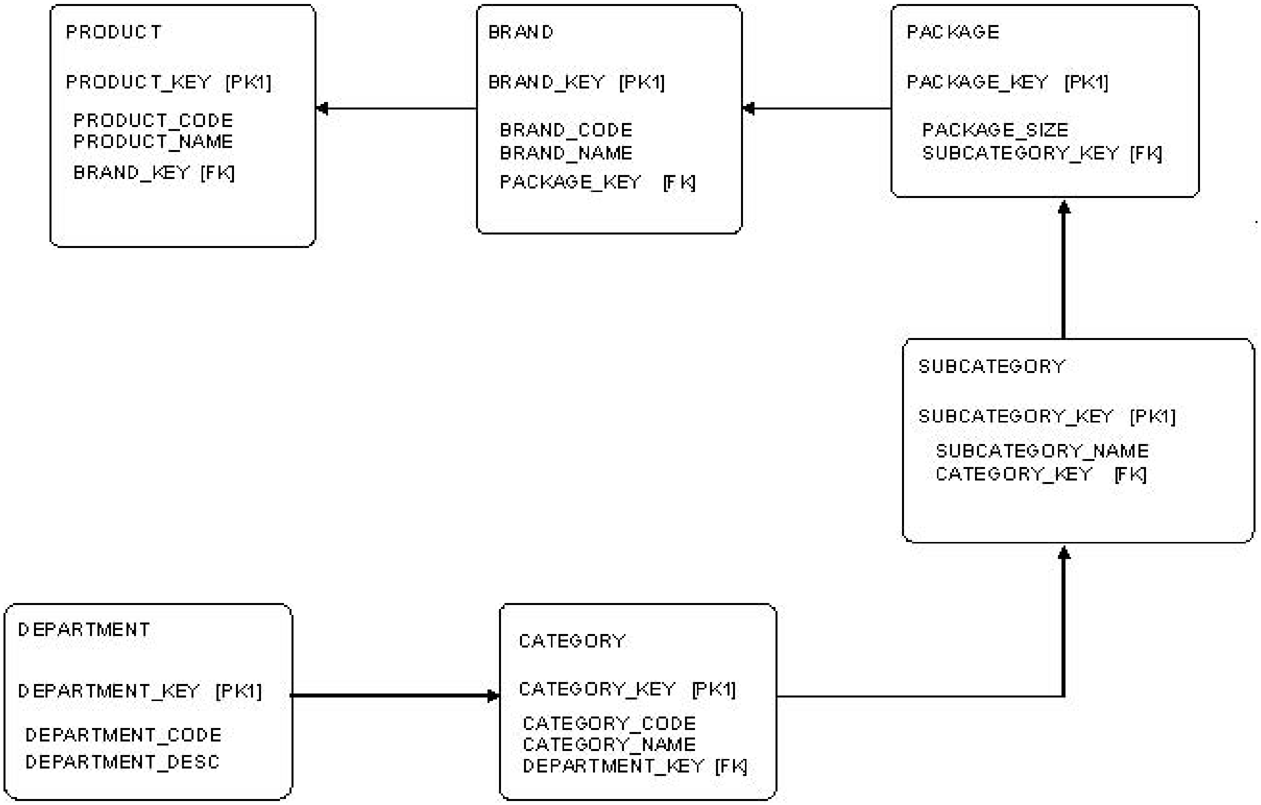


Figure 5: Snowflake product dimension.

Relational (Normalized) Schemas

The relational schema (normalized data structures) is the third data model used in the data warehouse environment. It is built using the modeling technique normalization.

Normalization splits tables into multiple tables to eliminate redundancies and possible update anomalies. The concept here is that there should be a single (well, as few as possible) instance of a value. For example, let us say you have two tables—orders and customers. If there are a lot of orders stored for a single customer and the customer moves, then changes in the customer address column(s) in the customer table is all that is needed. By updating the address at only one place gives the system less chance of data anomaly. If you were to store the customer information in the order, then you would have to change all of the orders associated with that customer. While this is a simplistic example, it applies to complex situations. The relational (normalized) schema is commonly used in the enterprise data warehouse where the user community queries the data from the data marts.

Defining the Metadata World

Metadata is a critical piece to tie all the pieces of the data warehouse/data mart process. It is the central repository of data about data. Metadata is the blueprint of how the data warehouse was built and what it means. Just as there are different types of blueprints, there are different types of metadata. They are (1) business metadata, (2) technical metadata, and (3) contextual metadata.

Business Metadata

Business metadata describes the data from an organizational perspective. It is written in the terms that the organization understands. Business metadata won't contain transformation rules like "Projected Profits = Column ASP039a × 0.67 – column ASP016c." It would contain "Projected Profits equals Projected Income times Percentage after GM&A Costs minus Projected Costs." Business metadata should contain items like

Confidence in data;

Confidence in data quality;

Who is responsible for the data (data steward);

How the data was calculated;

Definition of terms; and

Valid values.

Technical Metadata

Technical metadata describes the data as required by the IT group to understand how the data warehousing environment was or is to be built. Technical metadata contains items like

What is the system of record for each piece of data?

What are the columns extracted from the source systems?

What are the rules for filtering the data out of the extraction?

What are the transformation rules that convert and summarize the data into the data warehouse?

What are the fields in the data warehouse and what are their meanings?

What were the data reconciled with and what are the values?

Contextual Metadata

Contextual metadata is the hardest metadata to collect, but some of the most important metadata there is. Contextual metadata comes from systems outside of your normal transaction system and helps put context around the data. If you were running a campaign for soft drinks during the July 4th holiday (U.S. Independence Day during the heat of the summer), you might look to see how sales were during the past five campaigns. Let us say that 1995 had $1.0 million in sales, 1996 had $1.1 million, 1997 had $0.5 million, 1998 had $1.25 million, and 1999 had $1.5 million. Contextual metadata might keep track of the weather at each store location and when you drill down, you may find that in 1997 it rained. All of the 4th of July celebrations were called off and so no one bought soft drinks.

Contextual metadata might contain

Stock market trends;

Weather reports;

Political issues;

Religious issues;

Court actions; and

Headlines of the day.

Most organizations make metadata a "nice to have" requirement because they do not wish to take the time and effort required to originally build the metadata repository and keep it functional. However, metadata is critical to the total success of the data warehousing organization. Why? Because it is the only place that documents the true business rules as well as changes to these rules over time.

Several factors as to why metadata is needed in the data warehouse environment are (Marco, 2000)

Current systems are inflexible and nonintegrated;

The existing data warehouse and data marts need to grow;

Business users' needs are not being fulfilled;

Organizations need to reduce the impact of employee turnover; and

Organizations need to increase user confidence in data.

Without a robust, architected metadata environment, these issues will have a hard time being met.

The Internet provides the best solution for the presentation of metadata. As more organizations provide more flexibility in work environments (work at home, VPN via Internet, etc.), using the same tool (the browser or other Web-enabled technologies) and the same techniques (search, point, and click) that the user community uses when accessing the data warehouse or data mart continues to drive metadata to be accessed via the Internet.

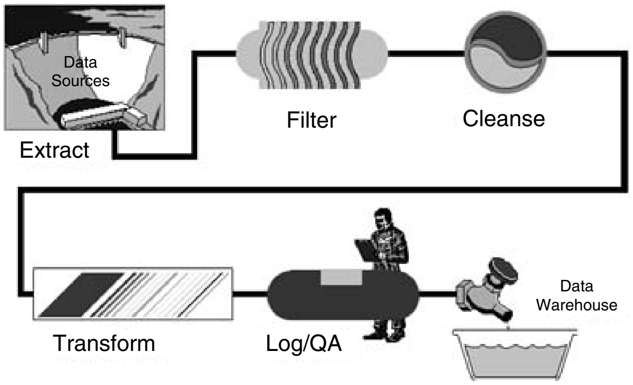


Figure 6: Information sourcing window.

The Extraction, Transformation, and Load (ETL) Processes

The extraction, transformation, and load (ETL) processes are needed to move the data from the operational systems, whether they are order entry, reservations, or Internet clickstream data, into the data warehouse and from the data warehouse to the data marts. The information sourcing workflow is broken into five processes (Figure 6):

Extract—This process takes data from the operational systems. The output of this process is raw data.

Filter—This process takes the raw data and discards "noise" data from the data set. The output of this process is dirty data.

Cleanse—This process takes the dirty data and checks the data quality and corrects the data, if possible. The output of this process is clean, high-quality data.

Transform—This process takes the clean data and restructures and summarizes the data. The output of this process is clean, consistent, summarized, and useful data.

Log/QA/Metadata—This process takes the useful data and does a final check to make sure the data passes the quality assurance test (reconciliation) and produces the metadata for the workflow. The output of this process is verified data and metadata.

Extraction

Extraction is the pulling (or pushing) of the data from the operational systems to the ETL process. The goal of the extraction component is to take only the necessary data from the operational systems and prepare it to go through the rest of the ETL process.

In order to do this, the system of record (best source of data) needs to be determined. The system of record is the best data for this part of the data warehouse. For data warehouses whose goal is to understand sales in a store, determination must be made as to which data best represent the sales data. This could be cash register receipts or it could be inventory brought in. The system of record is determined by looking at all of the sources of the data and choosing the most accurate. The system of record of store sales could be the source of data coming into the store versus the source of data being sold from the store. The output from the extraction process is a subset of raw data that will be processed.

Filter

Filtering the data takes the raw data and discards the "noise"—purely operational—data. For example, date_last_printed is probably not something that the data warehouse user community would want to do analysis on. Neither would flags and switches be useful in the data warehouse. The result of the filtering process is dirty data, but data that will most likely be needed later in the process.

Cleanse

Once the data have been filtered, there needs to be some cleansing to make that dirty data clean. Cleansing the data is important so that you won't have lots of duplicates that seem good, but really aren't. Customer names and addresses always need cleansing. If a query requests the top five customers and the results showed

AT&T
A T & T
A T&T
International Business Machines
IBM

there would be a high degree of skepticism about the validity of the data.

Data quality is extremely important in the data warehouse due to the fact the skepticism about the validity of the data in the data warehouse will kill the usefulness of the data warehouse environment. Data quality is more than cleaning names and addresses, although that is a large component of it. It is also about making the data consistent across multiple systems. Issues like single code definitions and consistent abbreviations for consistent retrievals across different source systems will help an organization with its analysis.

There are a number of products and services in the data quality marketplace. Some organizations are exploiting the Web by providing cleansing tools accessed via the Internet using an application programming interface (API). This allows for the cleaning and merging of the data into high-quality output without purchasing the product.

Poor data quality can undermine the organizations ability to focus on the customer efficiently. For example, poor data quality in a customer relationship management (CRM) system might

Cause the inability to track a customer through all facets of your supply chain;

Not allow the cross- or up-selling of the customer; or

Increase returned e-mail or paper mail (snail-mail), which increase the costs of marketing and hurt customer satisfaction.

However, a return on investment (ROI) study should be made on the data quality process to make sure that the implementation meets the needs of the organization without too high a cost.

The output of the cleansing process is clean, high-quality data.

Transform

Once the data are clean, there needs to be a process that transforms the data into the new data structures as well as summarizes the data. For example, if keeping the number of pay-per-view movies each customer watched during a day was the subject area being built, then there would be a single row in the fact table showing a customer watched three movies during that day. There may not be a need for a finer level of granularity, e.g., a separate row for each movie—unless that someone was running for public office! The output of the transform process is clean, consistent, and thus useful data.

Log, Quality Assurance Checks, and Metadata

The last step in the process is to take the useful data and run some quality assurance checks and to make sure the data reconcile with the source data. There are multiple ways to do this. Some of them are (1) check totals against the source file, (2) compare the totals to a report generated from the source system, and (3) compare the results to a previous day's data to make sure it is in the ballpark (within a certain percentage, for example). Once that has been accomplished, log that fact in the metadata and publish the data.

The Hardware Architecture

The server that houses the database for the data warehouse and data marts needs to provide for a completely different type of computer environment than the transaction/OLTP systems. There are three computer architectures that could be used in the data warehousing environment. They are

Symmetrical multiprocessing (SMP) architecture (Figure 7);

Clustered architecture (Figure 8); and

Massively parallel processing (MPP) architecture (Figure 9).

Symmetrical Multiprocessing (SMP) Architecture

Symmetrical multiprocessing (SMP) architecture (see Figure 7) is a share-everything architecture. The memory and disk are accessible to all the central processing units (CPUs) in the box. This provides for a scaleable environment, to some point, so that additional CPUs will increase the performance of the data warehousing environment. SMP is proven technology and has been around for quite awhile. Most general-purpose database systems (DBMSs) work well in the SMP environment. A variation on SMP is the nonuniform memory access (NUMA) style of SMP. While SMP accesses memory uniformly across all processors, NUMA promises increased performance by having non-uniform memory access. For example, a two-processor NUMA machine may have one processor accessing 60% of the memory and the other 40% of the memory. While the architecture looks promising, there are not many implementations of NUMA to date.

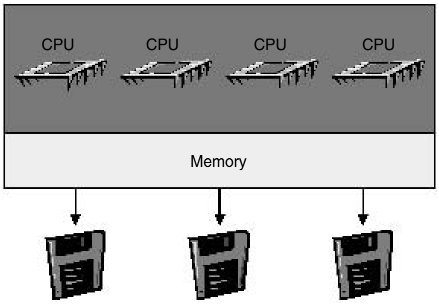


Figure 7: Symmetrical multiprocessing (SMP) architecture.

Clustered Architecture

Clustered architecture (see Figure 8) is a group of systems (single CPU or SMP) loosely coupled together with a high-speed interconnect. Each system has its own memory but is able to share the disks. Clustered architectures provide strong failover capabilities, which are good for transaction systems, but of less value for the data warehouse or data mart. Most general-purpose DBMSs work in a clustered architecture, but have many limitations for performance. Real-time data warehouses make use of the clustered architecture.

Massively Parallel Processing (MPP) Architecture

The massively parallel processing (MPP) architecture (see Figure 9) is a share-nothing architecture. It does not have

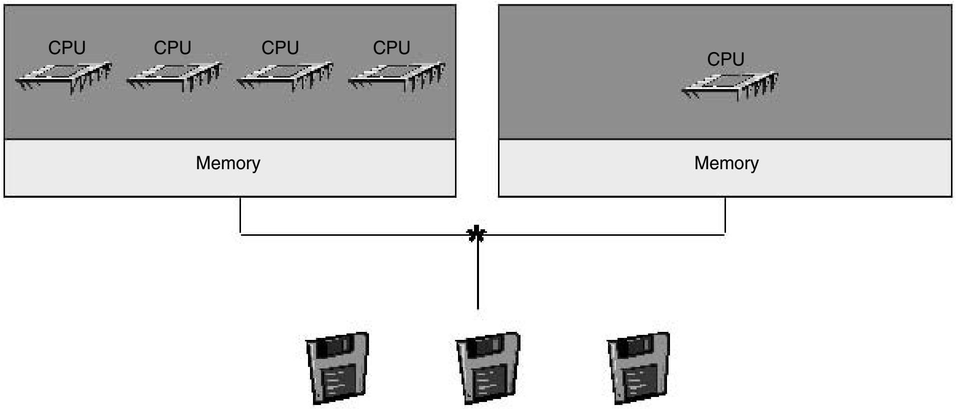


Figure 8: Clustered architectures.

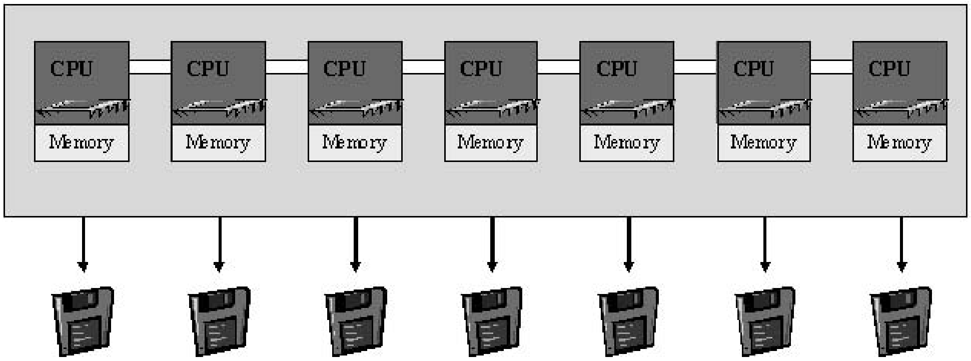


Figure 9: Massively parallel processing (MPP) architecture.

sharing of memory or disks. However, there is a high-speed interconnect that ties the CPUs together. MPP architectures are perceived to be highly scaleable. However, few proven data warehouse implementations are available. Most of the DBMSs do not (yet) take advantage of the MPP environment. Teradata, one data warehouse DBMS, actually is written to run like a MPP environment. It is done in software instead of hardware or in a combination of software and hardware.

The End-User Access World

Internet query tools are quickly replacing the first generation of desktop tools in the Decision Support System/Executive Information System (DSS/EIS) world. Accessing the data requires three components—(1) connecting to the appropriate data source, (2) executing queries against the data, and (3) viewing the results in a meaningful way.

Most tools for accessing the data warehouse in the market today are Web-enabled. This allows the data to be shared across the whole organization. This sharing of the information allows for highly interactive, collaborative analysis that helps organizations keep costs down. This sharing produces a better targeting of customer that can lead to higher retention of the customer base. Marketing knows that it is cheaper to retain a customer than to get a new one.

The user community should help in choosing the products that will solve their needs. IT should never try to force products on the user community. Some of the pitfalls when choosing an end-user tool include

Failure to ease the burden of data identification, definition, and access method;

Failure to adequately train the end-user community and IT;

Failure to provide data in a timely fashion; and

Failure to understand the users' requirements for data access.

To be successful, IT needs to ensure the data access tools and components work together seamlessly. This may mean building a data mart for the tool. If the tool works best in a snowflake schema, then build a data mart using the snowflake schema.

Tools for the user community are plentiful. The tools come with different styles of data manipulation and presentation:

Managed query;

Integrated query and analysis using OLAP operations (such as slicing, dicing, roll-up, drill-down, drill-across);

Visualization techniques; and

Data mining.

Each of these styles may be needed for your user community to be effective. OLAP and data mining are covered elsewhere in *The Internet Encyclopedia*. Today, these styles are done primarily in the tools used, but a strong trend in database technology is the extending of the functionality of transaction databases to include OLAP and data mining operations.

Another important task for a successful user community is that IT needs to set user expectations and monitor them during data access. Remember, returning 10 million rows back to the desktop in 10 seconds is not feasible. Depending on your network, it could take an hour or more.

The Internet and the tools available with an Internet interface are going to define the user perception of the data warehousing environment. Care needs to be given to make sure that they fit in the larger organizational strategy.

Some data warehouse activity cannot get true user feedback unless the organization specifically asks for it or the customer specifically gives it. These organizations can still use data warehousing to provide much value to the end-user. We can see examples of this in many places. On the Internet, go to http://www.Amazon.com and search for your favorite author or the title of your favorite book. Quickly, it gives you a list of other books by the same or by a different author on the same subject that others have purchased that might be of interest to you. Some quick market basket analysis has been done and the results of that were provided to you. Now with the Internet and clickstream data, you can study the purchasing habits of your customers to see whether any new patterns are starting to show up or whether you are benefiting from this new knowledge (market basket analysis). Measurement of the effectiveness of the new knowledge and the implementation of the new strategy is starting to become very precise.

THE CLICKSTREAM DATA WAREHOUSE

The Internet has had a major influence on processing today. The potential (and promise) of the Internet can be seen in understanding what a visitor or potential customer is doing at your Web site. Every click made can be captured and loaded into the data warehouse or data mart, unless you are barred from collecting this information by law as the Federal Government is. This is called clickstream data. The analysis of these data is called clickstream analysis.

The clickstream provides the ability to "watch" what the customer is doing. A Web site owner will be able to see what the customer is looking at and infer what they are not looking at. Then by looking at what the customer purchased, intelligence can be gathered about user perception. The Internet provides a wealth of knowledge for the user community over brick-and-mortar stores.

Some of the knowledge available from clickstream data is

What is the customer looking for?
What are they comparing it to (at your Web site)?
What did they purchase?
Was your site easy to navigate?
How long did they browse at your site?

The problem with clickstream data is that it must be looked at in chronological order (for the most part). There are only a few instances where looking at certain click instances can gather knowledge. Chronological order may mean capturing time down to the second, not just the day like most data warehousing environments.

A second problem deals with the voluminous amount of data associated with the series of clicks in a clickstream. There needs to be a process in place to capture, transform, and load the data for analysis, not only for immediate business needs but also for historic analysis. Both will be extremely useful for the user community.

The Internet, with this clickstream data, is a start at getting into the thoughts of the customer. Transaction data are not always trustworthy (although it seems we believe it wholeheartedly today!) to provide a picture of what is the preference of the customer. If Henry Ford had created a data warehouse early on in the days of automobile and did analysis on colors of cars that people bought, black-colored cars are all that would be available. All the transactions would have color = black, so why paint a car any other color? (*Note:* Henry Ford is credited to have said you can have any color automobile you want, as long as it is black.)

Clickstream data can deliver much business intelligence. Capturing the appropriate data will help to understand the reasons why customers buy or not. The Internet, with the clickstream data, is in its infancy state in capturing this data. Analysis on the data is being done mostly by the early adopters.

Another type of clickstream data is metadata clickstream data. Metadata is the central repository of data about data. Analyzing clickstream data about usage of metadata can help the organization understand how well the user community is able to find data within the data warehouse.

PRIVACY

Since so much data are being captured via the Internet and transaction systems, it will not be long before there is more sharing of data between organizations. Today, there is a desire among some customers for more privacy. Most organizations have an opt-out process, which allows the customer to notify the organization that it does not want the organization to provide the customer data to other organizations. This is being done through Web pages as well as interactive voice response systems. It is expected that governments will require organizations to have an opt-in process. The opt-in process allows the customer to notify organizations that the customer would like to have his or her data provided to other organizations.

CONCLUSION

The Internet plays an important role in the data warehouse and data mart environment. With the ability to access data from anywhere in the world with your Web browser, the user community is asking to see relevant data in a timely fashion using the Internet. Data warehousing has evolved due to the inability of the user community to gain access to the data. Today, an architected environment needs to be defined and built—regardless of whether you build a collection of data marts or a data warehouse feeding data marts. One thing is for sure—the data warehouse environment is *not* an archival system and *not* an audit system. The data warehousing environment *is* a tool for the analysis of data.

Building the data warehouse and data mart environment will take some time mostly due to the fact that the user community does not know what they want until they actually start seeing data. This goes against the grain of what is being taught in colleges and universities today about how to build systems.

However, for organizations to grow and become more prosperous, they need to be able to

capture data;
transform that data into information;
discover new knowledge from that information;
test that the new knowledge is true; and
deliver the new knowledge to the user community.

Clickstream data are providing organizations with data that helps get into the thought processes of their customers. While this analysis is still in its infancy, it is already providing new knowledge for the data warehouse. This will allow organizations to produce more targeted advertising campaigns, better focused one-to-one marketing, and stronger customer relationship management.

This is the *value* of the data warehouse.

GLOSSARY

Clickstream analysis Analysis of a visitor's clicking through a Web site.

Clickstream data The data collected as a visitor to a Web site "clicks" from page to page.

Database A system that provides for the storage and retrieval of data.

Data warehouse An environment that captures all of the data used for analysis for an organization—whether this environment is a monolithic database, distributed database, or a collection of data marts is determined by the requirements.

Data mart A platform for a subset of data warehouse data for analysis serving a single department, subject area, application, or limited part of the organization.

Knowledge The learning of something new to help an organization provide better products and services.

Normalization A data-modeling technique for transaction systems that breaks apart tables to minimize redundancy and to alleviate update anomalies, with the most common degree being third normal form.

OLAP (online analytical processing) A type of computer processing that enables the user community to analyze data in a fast and consistent way, with typical operations including drill-down, roll-up, drill across, reach through, and rotation.

OLTP (online transaction processing) A type of computer processing that captures information and responds immediately (online) to user requests (transactions).

Snowflake schema Part of the dimensional model that normalizes (or breaks apart) dimensions of a star schema to a group of tables in order to alleviate redundancy and possibly update anomalies.

Source system A system, usually a transaction system, that feeds the data warehouse, consisting of external data, spreadsheets, and documents.

Star schema Part of a dimensional model, which has a central fact table surrounded by dimension tables (i.e., points of a star).

Strategic Dealing with the forwarding thinking of an organization as to where and how the organization should change.

Tactical Dealing with the day-to-day issues of the organization.

Transaction system A system that runs the day-to-day business (e.g., reservations system, accounting systems, manufacturing systems) in either an online or batch processing environment.

User community Any group of people who the data warehouse/data mart provides value to, which can include employees, vendors/suppliers, or customers.

CROSS REFERENCES

See *Data Mining in E-Commerce; Databases on the Web; Intelligent Agents; Machine Learning and Data Mining on the Web; On-Line Analytical Processing (OLAP); Personalization and Customization Technologies.*

REFERENCES

Inmon, W. H. (1992). *Building the data warehouse*(1st ed.). New York: QED Press.

Inmon, W. H., & Kelley, C. (1993). *RDB/VMS: Developing the data warehouse.* New York: QED Press.

Kelley, C. (2001, April 12). Is it knowledge or shared knowledge? *Data Warehousing and Business Intelligence.* Retrieved March 29, 2003, from http://searchdatabase.techtarget.com/tip/1,289483,sid13_gci765320,00.html

Kimball, R., Reeves, L., Ross, M., & Thornthwaite, W. (1998). *The data warehouse lifecycle toolkit.* New York: Wiley.

Marco, D. (2000). *Building and managing the meta data repository.* New York: Wiley.

FURTHER READING

Adelman, S., Bishchoff, J., Dyché, J., et al. (2002). *Impossible data warehouse situations: Solutions for the experts*. Reading, MA: Addison-Wesley.

Adelman, S., & Moss, L. T. (2000). *Data warehouse project management*. Reading, MA: Addison-Wesley.

Barquin, R., & Edelstein, H. (1997). *Building, using, and managing the data warehouse.* Englewood Cliffs, NJ: Prentice Hall.

Inmon, W. H., Imhoff, C., & Sousa, R. (1997). *Corporate information factory*. New York: Wiley.

Kachur, R. J. (2000). *Data warehouse management book*. Englewood Cliffs, NJ: Prentice Hall.

Kimball, R. (1996). *The data warehouse toolkit*. New York: Wiley.

Kimball, R. (2000, January 20). The special dimensions of the clickstream. *Intelligent Enterprise Magazine*. Retrieved March 29, 2003, from http://www.intelligententerprise.com/000120/webhouse.shtml

Kimball, R., & Merz, R. (2000). *The data Webhouse toolkit: Building the Web-enabled data warehouse*. New York: Wiley.

Silverston, L., Inmon, W. H., & Graziano, K. (1996). *Data model resource book*. New York: Wiley.

Tannenbaum, A. (2001). Metadata solutions: Using metamodels, repositories, XML, and enterprise portals to generate information on demand. Reading, MA: Addison-Wesley.

Denial of Service Attacks

E. Eugene Schultz, *University of California-Berkeley Lab*

WHAT ARE DoS ATTACKS?

Background

News about some kind of disruption or prolonged outage of computing services due to malicious activity or programs seems to surface almost every day. Consider, for example, how in late 2001 a flood of network traffic brought the *New York Times* network to a standstill. Earlier that year, the Web server of the Computer Emergency Response Team Coordination Center (CERT/CC) was brought down by a denial of service (DoS) attack. A series of DoS attacks in February 2000 brought down numerous systems used by ZDnet, eTrade, Amazon.com, eBay, and others, severely disrupting these companies' ability to conduct ebusiness transactions. Other large companies that have been successfully targeted for denial of service attacks include Microsoft and Yahoo!

Although the public may not know exactly what DoS attacks are, the DoS attacks certainly know something about them. CloudNine, an Internet service provider (ISP) recently went out of business. Several media sources attributed CloudNine's business failure to a series of DoS attacks that this ISP encountered. About the same time that CloudNine went out of business, the University of Colorado shut down what was then the Internet's oldest Internet relay chat server because of repeated denial of service attacks. A recent study conducted at the University of California at San Diego indicated that approximately 4000 DoS attacks occur every week. (See Lemos, 2001, available at http://news.com.com/2100-1001-258093.html?legacy=cnet).

DoS attacks have grown from a relatively rare phenomenon to a frequent occurrence in recent times. Whereas 10 or 15 years ago DoS attacks were relatively unheard of, the relative proportion of DoS attacks has in recent years increased dramatically. CERT/CC has repeatedly stated that more DoS attacks are reported to this organization than any other type of attack.

First we'll turn our attention on how DoS attacks can be distinguished from other types of undesirable events and other types of attacks.

Distinguishing DoS Attacks

DoS Attacks versus Other Damaging and Disruptive Events

What distinguishes DoS from other damaging and disruptive events? DoS attacks are designed to disrupt, overwhelm, and/or damage computing resources and data. The fact that they are *designed* to cause disruption, damage, and so forth implies *intention* on the part of the attacker(s). Many disruptive and damaging events occur in normal computing environments. Examples include allowing disks to fill up, running software with bugs that cause the software to crash frequently, misconfiguring routers, firewalls, and Domain Name System (DNS) servers, faulty system configurations that result in floods of broadcasts or other undesirable events, and faulty network configurations that result in poor traffic flow. These events do not normally constitute DoS attacks, however, because they have not been deliberate. Human error is in fact a far greater cause of loss in the world of computing than intentional attacks (Schultz & Shumway, 2001). Still, DoS attacks are a major (and ever growing) source of concern.

DoS Attacks versus Other Types of Attacks

DoS attacks are at least to some degree different from other types of attacks. Information security has three widely recognized goals—confidentiality of data, integrity of data and systems, and availability of data, systems, networks, and services. (Many experts feel that focusing exclusively on these goals represents too narrow a view of the goals of information security, however. They assert that other goals such as nonrepudiation of actions and transactions, i.e., being unable to deny that one has initiated an action, business order, and so forth, and accountability, i.e., having records of each user's actions performed on computing systems and networks, are also among the major goals of information security.) DoS attacks target primarily availability. To this end, DoS attacks can be distinguished from attacks geared toward copying or stealing confidential data or tainting the integrity of data and/or systems. DoS attacks can, however, overlap with attacks on integrity—someone who breaks into a system and changes key parameters in critical system files to out-of-range values is in effect attacking not only the integrity of the system, but also its availability. Depending on the particular changes made, the victim system might not only crash, but may not even be bootable—something that is likely to constitute a serious loss of availability over a prolonged period of time.

Motivations for Launching DoS Attacks

Why do attackers launch DoS attacks? Although we cannot usually be sure of the exact motives of anyone who launches a DoS attack, case studies in which DoS attacks' origins have been traced to individuals indicate that in many cases DoS attacks are launched with one or more distinct motives:

Retribution, Hostility, and/or Frustration

The perpetrator may be an employee or ex-employee of a corporation who perceives that s/he has not been treated fairly. The perpetrator may accordingly try to get even with the person or organization perceived as causing the inequity. In one case a database programmer wiped out the entire customer database of an insurance company in retaliation for his employment having been terminated. In another more recent case Timothy Lloyd was convicted and sentenced to 41 months in jail for planting a "time bomb" in the network of his employer, Omega Engineering, causing 12 million dollars in damages. (A time bomb is a destructive program that activates when the system clock reaches a certain time; see Gaudin, 2002, available at http://www. nwfusion.com/news/2002/0304lloyd.html.) Lloyd apparently planted the time bomb because of growing frustration due to what he perceived as diminishing ability to influence his work environment.

Need to Gain Recognition or Regain Lost Status

Sometimes perpetrators launch DoS attacks, usually in the form of some kind of sabotage, then rush in and solve the problem they have introduced quickly in the hope of being recognized as a "hero." Because they have created the problem, they can quickly solve it. Having quickly solved it makes them appear to have incredible abilities.

Political Activism

In some cases the perpetrator of a DoS attack may be politically motivated; the DoS attack(s) may represent an attempt to register a protest against the actions and/or policies of a particular organization or government. In 1989, for example, a perpetrator launched a worm called "WANK" (Worms against Nuclear Killers) to protest the U.S. National Space Aeronautics Administration's (NASA's) impending launch of a rocket containing a nuclear-powered energy generator. The terrorist attacks in the U.S. on September 11, 2001 also convince many information security experts that widespread attacks on computers and networks can and will be launched by organizations such as the one reportedly responsible for the September 11 attacks.

Gaining Control over Computing Resources

Perpetrators of DoS attacks have sometimes launched attacks to monopolize computing resources. Chat channel users (often but not always members of the "hacker" community) have, for instance, initiated DoS attacks to take over chat channels. Owners of certain Web sites have reportedly brought down competing Web sites for extended periods.

Attacking for the Sake of Attacking

Yet another motivation for initiating DoS attacks has been to bring systems and/or services down simply for the sake of doing so—a manifestation of "electronic vandalism."

"Fame and Glory"

Sometimes individuals have launched DoS attacks to gain notoriety—to bring them attention (often thereby elevating their status in "hacker circles").

Information Warfare and/or Economic Espionage

In a number of cases individuals have initiated DoS attacks as part of an apparent information warfare or industrial espionage effort. When the U.S. military bombed Yugoslavia in 1999, for example, individuals within Yugoslavia countered by launching flooding attacks against U.S. military systems in the Pentagon. A similar scenario occurred in 2001 after a U.S. spy plane crash-landed within the territory of the People's Republic of China (PRC). For weeks DoS and other attacks on U.S. computing systems originated from IP addresses within the PRC.

Now that we have explored the reasons for perpetrating DoS attacks, let's next focus on why attempted DoS attacks have such a high probability of succeeding.

Why DoS Attacks Succeed

DoS attacks are not only frequent, but the probability that they will succeed is high. Why do DoS attacks succeed as often and well as they do? Let's consider several explanations.

Internet Design

The Internet itself has not been designed with security in mind. The Internet was originally built for providing free and easy access, not for assuring confidentiality of data, integrity of data and systems, continuous availability of services, and so forth. It is safe to say that few if any of those who originally designed and implemented the Internet ever imagined that people would deliberately try to interrupt or corrupt services delivered over the Internet, bring down critical network devices, and so on. If you want security on the Internet, you have to provide it yourself.

Many if not most protocols and services do not incorporate protections against DoS attacks. If you look at the RFCs (Requests for Comments—see http://www.ietf.org) for networking protocols and services, you'll see that many of them address security concerns, but that few of them include mechanisms for staving off DoS attacks. Patches that fix vulnerabilities in major implementations of these protocols and services often have been developed as an afterthought, but add-ons seldom address software problems as effectively as addressing requirements during the initial development cycle. One of the best examples of a protocol that offers little protection against DoS attacks is the UDP (User Datagram Protocol) protocol. UDP incorporates virtually no reliability mechanisms; it is sessionless, connectionless, and "unreliable" (although extremely conducive to good network performance). It is often easy to simply flood another system with UDP traffic, causing a certain service or often the system itself to hang or crash. Any service (such as the Network File System or NFS in Unix and Linux, and Simple Network Management Protocol or SNMP) based on UDP (as well as other protocols) is thus a prime candidate for DoS as well as other types of attacks.

Bug-Ridden Software

Numerous vulnerabilities in many software products have been identified over the years. Many of these vulnerabilities can be exploited to produce DoS conditions. Buffer overflow attacks, to be covered shortly, provide one of the best examples. Worse yet, certain vendors' products, most notably Microsoft's, are interrelated—failure in the functionality of one can result in massive failure. Their products are deployed so much that they comprise what functionally constitutes an "information technology monoculture." When there is a DoS susceptibility problem in one product, many related products have an elevated susceptibility to DoS attacks, leading to the potential for widespread havoc and disruption.

Increased Popularity of DoS Attacks within the "Hacking Community"

Whereas break-ins were in many respects the modus operandi within the "hacking community" several years ago, DoS attacks are ostensibly often now more valued within this community. With a single DoS attack an attacker can cause considerable trouble, thus gaining considerable notoriety within this community. Consider the recognition that "Mafia Boy," a Canadian teenager, gained after he reportedly initiated successful DoS attacks against several companies that made front-page news in February 2000.

Slowness in Recognizing the Problem

Computing itself is still to a certain degree unreliable. It is easy to confuse unreliability with susceptibility to DoS attacks. Additionally, for many years people did not take DoS threats very seriously because other problems such as break-ins into systems caused more loss and disruption at that time.

Corporations are not deploying necessary countermeasures. A report written by the National Academy of Science's Computer Science and Telecommunications Board (CSTB) asserts that U.S. corporations are not deploying security countermeasures needed to defend their computing assets from cyberattacks (National Academy of Sciences, 2002). This failure to adopt necessary measures also elevates the likelihood of successful DoS attacks.

We'll now consider the amount of damage and disruption that DoS attacks cause or potentially *can* cause.

The Toll

Losses associated with any type of information security attack are difficult to pinpoint with any precision. Consider that one estimate of the collective losses due to the distributed DoS attacks (to be covered shortly) against several U.S. corporations in February 2000 was five billion dollars. Was this estimate believable? No one really knows. Undoubtedly, the victim companies experienced a substantial level of financial loss—one that at least to some degree adversely affected each company's "bottom line." DoS-related losses are, however, generally attributable to a number of sources, including the following:

Down time in business-related or other critical systems. This source of loss constitutes an operational disruption that generally has the worst impact in billing and other types of financial systems in which down time of only a few minutes can be financially costly. Consider also the impact upon safety of a DoS attack on a flight control or plant process control system. Disruption of ongoing computing operations not only means that personnel time will be wasted, but also that somehow the down time will be made up in a manner that does not conflict with ongoing computing operations. This is often a far greater problem than anyone other than operations specialists and senior level management can readily understand.

The cost. The cost of investigating and repairing the problem adds quickly. Manpower, particularly time devoted by technical staff, is by no means cheap. Finding the cause of a successful DoS and correcting the problem can involve a significant manpower investment. Additionally, because it is frequently so difficult to trace the real origin of DoS attacks, investigations of the cause of DoS attacks are often more lengthy and complicated than for other types of attacks. (In many DoS attacks the source (origination) address in packets sent to victim systems is spoofed (falsified), making tracing the real origin of the attack more difficult. "Spoofing" means using an IP source address other than the real address of the machine from which this kind of attack is launched.)

Legal costs. Although so far legal costs associated with DoS attacks have been relatively rare so far, the fact that eTrade was sued by a customer who complained that he suffered stock market losses as a result of the DDoS attack against this company is significant. The cost of prosecuting suspected perpetrators of DoS attacks could also add to the financial toll of such attacks.

Loss of customer or third-party partner confidence and reputation. Disruption of access to services that customers and/or third-party business affiliates experience negatively impact confidence in the service supplier or other organization that experiences a DoS attack.

Interestingly, of the three major goals of information security, confidentiality, integrity, and availability, availability is in many organizations the most important. Confidentiality breaches can allow valuable or private information fall into the wrong hands, integrity compromises can cause people to doubt data and whether systems and network devices are working properly, but interruption of availability can bring organizations completely to their knees, so to speak.

DoS attacks are not likely to impact every type of site and organization equally. The next subsection discusses the types of sites that are most likely to be most negatively impacted by these attacks.

What Types of Sites Are Most Vulnerable?

Although predicting the magnitude of loss resulting from a DoS attack is by no means an exact science, some types of sites and computing environments are more likely to suffer damage and loss than others. The amount of probable loss depends on the types of damage and disruption (as discussed earlier) as well as factors such as the impact of down time upon an organization's business or mission. As mentioned previously, even a short down time in a billing system can have very adverse consequences for an institution. Someone from a stock brokerage firm that was one of the first to offer Internet-based transactions to customers told me that a certain newspaper reporter monitored that firm's Web site on a 24/7 basis and wrote a news article whenever the Web server was not functional. *As a general rule, the more critical continuity of services and operations are to an organization and the more mission critical the type of operational services, the more adverse the potential impact of DoS attacks is.*

The following are some representative types of sites that tend to be most vulnerable to DoS attacks:

Financial institutions and financial transaction processing centers

Stock and commodities trading centers

Military operations control centers

Transportation control centers (particular air traffic control centers, airports, and space flight centers)

Emergency response operations centers

Telecommunications operations centers

Plant process control centers

Security operations command centers

You may at this point have the impression that gigantic organizations with large, continuity-dependent operations centers are the only ones highly vulnerable to DoS attacks. Although it is true that risk due to the threat of DoS attacks is considerably elevated in such centers, DoS-related risk in small to medium-sized businesses is also often very high. Why? Because small to medium-sized businesses may operate with a minimum of monetary reserves or may face tight deadlines that affect the survival of the business. A DoS-related incident may be sufficient to disrupt operations to the point of forcing a smaller company out of business.

Now that we have considered the basics of DoS attacks and their impact, it is time to next consider the types of DoS attacks that can be found on the Internet and elsewhere.

TYPES OF DoS ATTACKS

To date numerous high-level types of DoS attacks have been identified. They may overlap with each other—they are by no means mutually exclusive of each other. Types of DoS attacks include the following.

Hardware and/or Software Sabotage

In a sabotage attack someone attempts to damage one or more hardware components and/or something related to the software within one or more systems. Normally a hardware attack requires physical access to hardware components. Someone who gains direct access to a system could, for example, damage the keyboard, video display terminal, any peripheral device such as an attached printer or mouse, or the CPU by deliberately dropping any of them. Alternatively, someone with physical access to a Unix or Linux machine may be able to boot the system in single-user mode, thereby gaining complete control over the system (often without having to even enter a password). Someone who cannot gain physical access can nevertheless damage system components such as hard drives by crashing the system suddenly. Another kind of sabotage attack involves integrity compromises in routers, firewalls, DNS servers, or ordinary systems (including applications that run on systems). An attacker can use methods such as unsecured Trivial File Transfer Protocol (TFTP) access to a router to put an altered access control list (ACL) that rejects incoming traffic or crafting bogus query replies that populate the DNS table of a DNS server with garbage, resulting in network failure. Critical system configuration files, binaries, databases, and so forth in virtually any system are also at risk of unauthorized alteration as a result of a wide variety of attacks, including breaking into a system, leaving the system dysfunctional, or causing it to crash.

Shut-Down or Slow-Down Attacks

Shut-down attacks are similar to sabotage attacks, except that the goal of a shut-down attack is not to cause damage, but rather simply to take one or more systems down. The attacker might power the system down, or remotely cause it to shut down by invoking a remote shut-down function. As mentioned previously, however, a nongraceful shut down can easily also result in damage to hard drives. Additionally, a perpetrator can slow down a system or network without shutting any system or device down. Although potentially less damaging to an organization's operational and business needs, a slow-down attack can nevertheless fulfill the intentions of someone bent on getting even with an organization or individuals.

Flooding Attacks

In flooding attacks someone or a program running on behalf of someone sends so many packets or data that they overwhelm the receiving system, service, or application. The best-known form of flooding attack is probably a SYN flooding attack in which the three-way handshake involved in establishing a TCP (Transmission Control Protocol) connection is misused. Normally in the first part of this handshake a system sends a SYN packet to another system. If a malicious system sends many SYN packets without responding to the target system's response, the target system may tie up system resources (e.g., memory) waiting for the malicious system's response.

System Resource Starvation Attacks

As mentioned previously, types of DoS attacks may overlap with each other. The previously mentioned SYN flooding attack is a flooding attack that results in resource starvation. Consider the effect of running the following routine on a Unix system:

```
while true
do
mkdir foo
chdir foo
done
```

In many Unix systems this will generate so many i-nodes (file system objects that hold file parameters such as ownerships and permissions, times of access and modification, and the size of each file) that the system may run out of resources and crash. A Windows NT attack tool named "CPU Hog" informs the victim system that it runs with a level of privilege unknown to the victim. The system feeds the process threads at the expense of other, unprivileged threads, causing the system to crash several seconds after this program starts to run.

Buffer Overflow Attacks

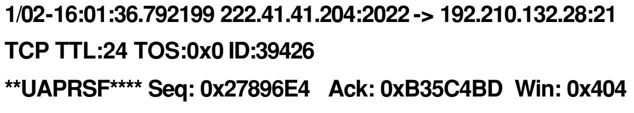
In a buffer overflow attack a program receives too much input for the amount of memory allocated. An attack can be as simple as sending a message with an excessively long subject line or using FTP (the file transfer protocol) to request a file with an excessively long name or to change directories, again with an excessive number of characters in the argument that follows the cd command. The excess input may be written into memory, resulting in the execution of rogue commands or programs, or it may simply cause the application (and possibly also the system shortly afterward) to crash because the application (and subsequently also possibly the system) goes into an abnormal running state.

Packet Fragmentation Attacks

Sometimes in the course of networking packets that are too large for routers or other network devices to handle are created and then sent over the network. Certain kinds of network devices, routers in particular, often cannot deal with such oversized packets. Packets are often consequently fragmented (broken into smaller pieces); each packet fragment will travel over the network until it reaches the destination system. If everything goes well, the destination system will then combine the fragment into one packet so that it can process the packet data appropriately. Numerous DoS attacks abuse packet fragmentation in one way or another such that the receiving system processes fragmented packets in a manner that causes it to hang or crash. One possible attack, for example, is to simply flood a system with fragmented packets. Another is to send fragments that overlap with each other in what is called a "Teardrop attack," where the overlapping packet fragments cause the reassembled packet to have values that are out of the permissible range. If the receiving system is not programmed to check and drop fragments constructed in this manner, it will go into an abnormal operating condition and crash. Yet other examples are the "Ping-of-Death" and "SSPing attack" in which a Windows 95/98/NT system is sent a series of fragmented IP packets. When the system combines the fragments, they now form a larger packet than the system can process, causing the system to freeze. "Jolt" and "Jolt2" attacks are similar, except that in these attacks multiple identical fragments are sent to the victim system (which in this case is a Windows 95/98/NT or 2000 system), causing massive CPU overutilization.

Malformed Packet Attacks

In malformed packet attacks the problem is that the program that implements a particular network service has not been written to detect and screen out malformed packets. RFCs define how packets for various networking protocols should be formed. Normal network programs form and then send packets that conform to these RFCs. An attacker can, however, use such a program to send packets with unconventional ("illegal") values, possibly that are out of range or are missing entire fields, to crash victim systems because they "spin out of control," possibly overallocating memory or using all available CPU. An example is a "Land attack" in which an attacker sends one or more packets that have the same source and destination addresses, something that confuses the receiving system to the point that it crashes. Another example is a "Bubonic attack" in which a barrage of psuedo-random TCP packets with identical TCP sequence numbers, source and destination ports, and other packet header information is sent to a victim Windows 2000 system and certain versions of Linux, causing it to crash. Still another example is a "Christmas Tree attack" in which the attacker sends a packet with every option set for a particular protocol being used. The most common type of Christmas Tree packet has all the TCP flags, SYN, URG, PUSH, FIN, etc., set. If the operating system of the target host has weaknesses in its TCP/IP stack implementation, the packet may crash or disable the system. Figure 1 shows a Christmas Tree packet captured by Snort, an intrusion detection tool, sent from IP address 222.41.41.204 to 192.210.132.28. Note in the beginning of the bottom line that all TCP flags, URG



```
1/02-16:01:36.792199 222.41.41.204:2022 -> 192.210.132.28:21
TCP TTL:24 TOS:0x0 ID:39426
**UAPRSF**** Seq: 0x27896E4   Ack: 0xB35C4BD  Win: 0x404
```

Figure 1: A Christmas Tree packet.

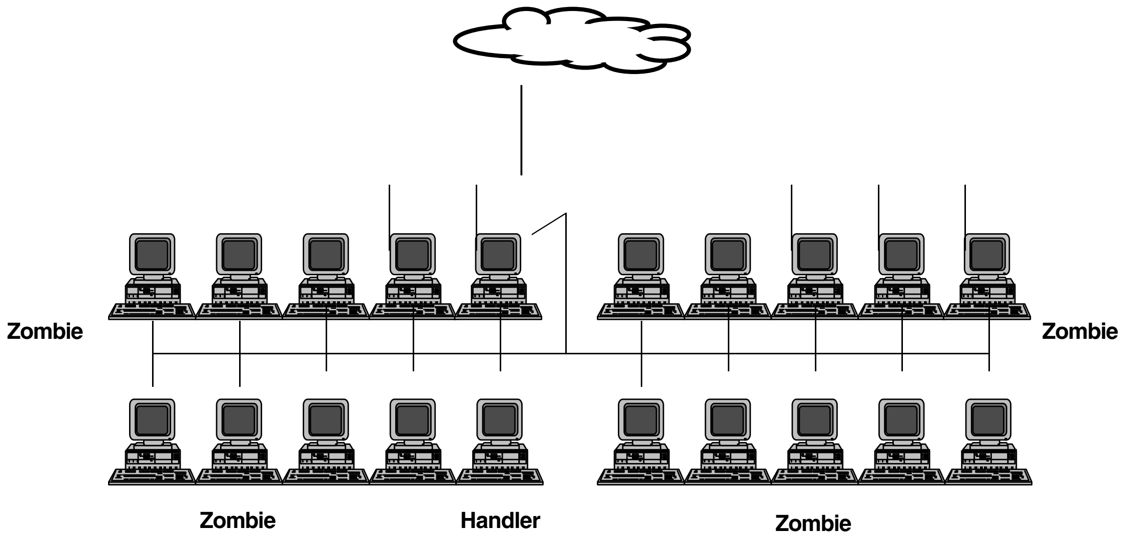


Figure 2: How zombies and a handler might be planted in a small network.

(U), ACK (A), PSH (P), RST (R), SYN (S), and FIN (F) are set.

"Boomerang" Attacks

In a "boomerang attack," the attacker spoofs the IP address of the intended victim host. In this kind of attack the attacker sends packets that elicit some kind of reply. If the attacker sends enough packets initially, a large volume of replies is returned to the victim, causing it to hang or crash. A representative type of boomerang attack is the "Smurf attack," in which the attacker sends many ping packets to systems within a network. A ping request in effect asks "Are you there?" Each pinged host replies, but each reply goes to the intended victim, overwhelming it.

Distributed Denial of Service (DDoS) Attacks

In DDoS attacks an attacker plants programs ("zombies") that unleash a DoS attack (for example, by releasing a huge volume of packets within a network or by releasing malformed or fragmented packets when they receive a signal from a specialized machine called the "handler"). Normally there are multiple handlers, one for each part of the victim network. Each handler receives its signal from a master system that starts the DDoS attack. The zombies and handlers are normally installed in systems well in advance of the intended date on which the DDoS attack is to occur. Figure 2 depicts a network in which zombies and a handler have been installed in a small network. The master is not shown here because the master is often outside of the victim network.

A set of conventions (e.g., special types of ping-like conventions) is typically used to keep the zombies and handlers and also the handlers and master in touch with each other before the attack is launched. This also enables the attacker to inventory how many zombies and handlers are in place and functional. In 2001 CERT/CC was the victim of a DDoS attack in which a vast number of large, fragmented UDP packets was sent to CERT/CC's Web port from numerous zombies. Web servers can usually deal with IP packet reassembly in a reasonable and efficient manner. Packet fragments must be stored until the servers receive every packet fragment; however; storing packet fragments in this manner consumes system resources. When too many packet fragments arrived, the Web server was overwhelmed, causing it to crash and resulting in a prolonged outage.

DDoS attacks constitute a particularly high level of threat because of the potential for massive outages. Nondistributed attacks are far less likely to cause a catastrophic condition—to reach a threshold necessary to disrupt traffic flow throughout an entire network, for example. Individual systems can do only so much, and networks and the systems therein have built-in mechanisms to correct errors and unusual conditions up to a certain point. So, for instance, when one machine is flooding another with packets, the latter is likely to send ICMP (Internet Control Message Protocol) "source quench" packets that suppress the former's packet sending. However, many systems, in this case, zombies, working in orchestration to attack the network are too much for normal correction mechanisms. In one case an attacker installed over one hundred zombies within an organization's network. When the zombies released a massive barrage of packets, the entire network went dead and could not be used for several days while technical people attempted to diagnose what was wrong and to bring the network back up. Many types of DDoS programs use stealth techniques to hide the presence of zombies and handlers, as well as to make the network communications between both unreadable through the use of encryption. Some recent types of DDoS programs even wait to create zombies—they create a process on each compromised system that later creates zombie programs, making detection of zombies even more difficult. The sheer number of DDoS tools (Trinoo, Stacheldracht, Tribe Flood Network (TFN) and its many variants, Shaft, and more) freely available on the Internet is another factor that elevates the level of threat from DDos attacks.

We've seen that there are many types of DoS attacks, each with potentially debilitating effects. Let's move on to the practical application of this chapter, namely how to prevent these kinds of attacks in the first place.

PREVENTION OF DoS ATTACKS

What can an organization do to prevent DoS attacks? This section addresses this topic by considering a number of alternatives. These alternatives are once again not necessarily mutually exclusive.

Risk Management Considerations

As mentioned earlier, for many reasons DoS attacks are likely to be successful. It follows that preventing DoS attacks altogether is no easy task. From a risk management perspective, therefore, a mitigation strategy that dictates prevention at all costs is not only likely to be unrealistic from a technical perspective, but also is likely to entail excessive cost in a cost–benefit analysis. A more realistic strategy is to determine the types of DoS attacks that could occur, the likelihood of occurrence of each, and the expected loss. Countermeasures that mitigate risk can then be deployed on a priority basis, with the risks that compromise the largest expected loss addressed before others until available resources are exhausted. So if, for example, extended downtime for billing systems due to sabotage or shut-down attacks were considered the greatest risk, priority in resource allocation could be devoted to controls that counter this threat. Classic risk management models would generally dictate this type of approach. Although this approach seems intuitive, it is at least to some degree flawed by inherent limitations of risk assessment methods. Simply put, these methods (especially quantitative risk assessment methods) imply a level of objectivity and precision that are simply not present. Furthermore, new instantiations of DoS attacks are constantly emerging, changing the relative risk of each identified source. Finally, although virtually every information security professional agrees that some form of risk analysis and risk management occurs in the course of conducting information security-related activity, the more formal the risk analysis and risk management method deployed, the more cumbersome and costly these methods are. By the time the output of these methods is available, the output is already likely to be out of date. In short, risk management-based methods of dealing with DoS threats (or any others, for that matter) are anything but a panacea.

What is the alternative? One of the most popular is the "due care" or "baseline controls" (or also what is somewhat arrogantly termed "best practices") approach in which an organization learns of the controls posture of other peer organizations and tries to deploy roughly the same controls posture. "Controls posture" means the type and amount of controls deployed. If, for example, other responsible peer organizations deploy a certain set of business continuity measures (as discussed shortly), an organization could exercise due care by deploying the same measures. One of the chief advantages of the due care approach is that organizations tend to adjust their controls posture based on real outcomes. If a financial organization does not deploy adequate perimeter control measures (e.g., firewalls), security-related costs due to intrusions, successful DoS attacks, and so forth can become intolerably high, prompting the organization to tighten its perimeter security. The due care approach tends also to be financially less costly; it can be based on normative data concerning actual controls deployment from various sectors, including the government, financial, transportation, and manufacturing sectors. On the other hand, many information security professionals are skeptical of this approach, which they often claim is too general to work in specific organizations that have specific business and security needs.

The debate between advocates of classic risk-based methods and the due care approach will continue. Notwithstanding, some kind of risk management activity is necessary to deal with DoS threats. At a minimum, organizations need to anticipate and prepare for worst-case DoS scenarios. Sadly, most organizations (i.e., those who experienced the massive DDoS attacks in February 2000) do not realize just how much disruption to ongoing operations and the financial impact of worst-case scenarios until after they occur. Consider, too, how little prepared most organizations currently are for a massive cyberterrorist attack designed to produce widespread disruption and panic.

Policy Considerations

Policy is in many respects the beginning point in countering DoS attacks. An organization's information security policy must at a minimum delineate baseline security measures needed to fend off DoS as well as other types of attacks. Policy provisions must also spell out actions (such as whether to bring in law enforcement) that are and are not to be taken in case of DoS attacks as well as punishments for employees, contractors, and others who initiate such attacks. Service providers (such as ISPs) and third-party business partners should establish clear expectations concerning the level of security to be provided for customers and business affiliates. One of the best ways to convey these expectations is through policy statements that can be communicated to employees through memorandums, Web site postings, e-mail messages, and so forth.

Business Contingency Measures

Given the vulnerability of today's systems, networks, and software to DoS attacks, it is naïve to assume that these attacks can be prevented altogether. Fallback measures, measures that enable an organization to carry on its business despite computing and network outages, are thus essential. Business continuity measures collectively provide one of the strongest sets of measures that minimize the impact of DoS attacks. Some of the major types of business contingency measures include the following.

Hot Sites

Hot sites are computing sites to which current computing operations can be rolled over in case of outage or disruption. The beauty of hot sites is that if set up properly, they provide an almost seamless method of operational continuity, although they tend to be rather financially costly.

Cold Sites

Cold sites are like hot sites, except that operations cannot be immediately rolled over to cold sites. Cold sites must first be brought into operational status.

Uninterruptable Power Supplies (UPSs)

UPSs are attached to computing systems and network devices in a manner such that if an electrical failure occurs, the computer or device keeps running because of availability of UPS-supplied power. The cost of UPSs has declined over the years to the point that UPSs are widely used in many organizations.

Redundant Array of Independent Drives (RAID)

RAID is a hardware solution that allows uninterrupted access to data on disks. A disk controller ensures that data are written to multiple disks. If one disk fails, data can be quickly accessed in one or more redundant disks.

Failover Systems and Devices

Failover systems and devices are redundant to those they support, such that if the primary system or device goes down, the failover one is immediately used. Failover systems are particularly important whenever there is a single point of failure, such as a firewall or router; single points of failure often exacerbate the impact of DoS attacks.

Firewalls

Firewalls are systems located between two networks that analyze and selectively handle (e.g., filter out, allow to pass through) traffic. Firewalls put up a barrier to attacks and undesirable types of network traffic. To the degree that DoS attacks can be prevented in the first place, properly configured firewalls are in many experts' minds the best single defense against this kind of attack. Although there are many different types of firewalls and various depths at which firewalls can analyze and handle traffic, the ability to filter traffic, i.e., allow it through or reject it, is the most basic firewalling function for stopping DoS attacks. Firewalls can, for example, block all incoming echo requests. One of the easiest types of DoS attacks to perpetrate is to send echo requests to another system. That system will reply with echoes that go right back to the original system, something that will trigger more replies, and so on, until both systems are flooded to the point of crashing. Using a firewall to block all incoming traffic destined for the echo port (TCP and UDP port 7) goes a long way in preventing echo storm attacks initiated from outside a network. They can also drop fragmented packets, probably the best measure against packet fragmentation attacks. Higher-ended firewalls such as application firewalls can even analyze input such as user commands to prevent commands (e.g., rm–rf *.* in Unix) that could cause DoS from reaching the intended victim(s). (rm means remove or delete. rm–rf *.* will recursively delete all directories and files, starting with the current directory, and without any prompt that allows confirmation of this drastic command.)

Routers

If properly configured, routers can serve as "low-level firewalls." They can filter packets according to their source and destination IP addresses, ports, and so forth. They can also drop fragmented packets. In so doing, routers can also help prevent certain types of DoS attacks.

Packet Filters and Personal Firewalls

Packet filters and personal firewalls are like miniature firewalls that run on systems to protect them from undesirable traffic. Packet filters are more basic—they allow or deny traffic based on rules such as the source of the traffic or the destination port. Personal firewalls generally go farther—they not only serve as packet filters, but they can also often block dangerous programs from being downloaded from Web sites, stop malformed packets from reaching a system, provide detailed log data, and so forth. One of the chief advantages of both is that they work regardless of whether an attack has been initiated from outside or inside a network. If configured properly, they (like firewalls and screening routers) block traffic that if allowed through would result in DoS.

Limiting Services

Many network services have vulnerabilities that leave the systems on which these services run vulnerable to DoS attacks. The recently mentioned echo service is one such service. The SNMP service is another. An attacker may be able to flood a system with SNMP traffic, causing it to crash. Other services that have proven especially conducive to DoS attacks include the character generator (chargen), WINS (the Windows Internet Name Service), POP (the Post Office Protocol), RPC (the Remote Procedure Call), the NetBIOS session service in Windows systems, and IMAP (the Internet Message Application Protocol). Running only the services necessary for business and operational needs is a good way of limiting the potential for successful service-related DoS attacks.

Patch Installation

Many vulnerabilities that if exploited can result in DoS have surfaced over the years. Attackers generally begin attacking by remotely scanning systems for vulnerabilities, then follow up with attacks that exploit the vulnerabilities they have discovered.

Most operating system and applications vendors produce patches, software, and other solutions that correct the vulnerabilities. Promptly installing patches is thus one of the best ways to prevent DoS attacks.

Quality of Service (QoS) Mechanisms

QoS (not to be confused with DoS) mechanisms have been developed to improve the quality of networking. These mechanisms include (among other things) a range of architectures, protocols, and routing mechanisms to boost performance and reliability. One of the side effects of the use of QoS is to make networks less vulnerable to DoS attacks. QoS mechanisms thus provide another defense against DoS attacks.

Intrusion Detection

Intrusion detection in simple terms is the process of discovering unauthorized use of computers and networks. Intrusion detection systems (IDSs) are somewhat limited in dealing with DoS attacks because they are post hoc in nature; they tell you that something is wrong only *after* something bad occurs. Nevertheless IDSs can be useful in helping prevent DoS attacks in that once they detect a DoS attack against one system, the data about the origin of the attack can be used to block ("shun") the IP address of the apparent attacking system at the entrance to a network (i.e., by modifying access control lists in the

Table 1 Preventative Measures against DoS Attacks

Preventative Measure(s)	Function or Mechanism	Effectiveness
Business continuity measures	Varied (ranges from simple hardware solutions to an entire hot or cold site)	High
Firewalls	Varied, ranging from packet filtering to connection analysis and management	High
Routers	Packet filtering	Medium to high
Packet filters and personal firewalls	Packet filtering and logging (plus additional functions in personal firewalls)	High
Patch installation	Correcting vulnerabilities	High
QoS mechanisms	Varied (throughput regulation, allocating priorities to network applications, etc.)	Medium
Intrusion detection	Detecting attacks	Low to medium
Third-party software	Varied (virus and worm detection and eradication, integrity checking, etc.)	Variable (depends on type and quality of software)

router and/or firewall), thus helping prevent further attacks of this nature.

Third-Party Software Tools

Finally, a variety of third-party software tools may potentially prevent at least some kinds of DoS attacks. The previously mentioned packet filters and personal firewalls are readily recognizable examples of such tools, but there are many others, too. Anti-virus software can, for example, help prevent DoS by detecting and eradicating self-reproducing programs such as viruses as well as many Trojan horse programs that users would otherwise download into their systems. Special software such as vulnerability scanners can detect DDoS zombies and handlers through methods such as determining whether certain ports known to be used by DDoS tools are listening for input. For example, many versions of Trinoo utilize UDP port 27444 by default. If port 27444 on a system is listening, a Trinoo zombie may be installed on that system. (Note that the netstat command can also list listening ports on a system.) Other software uses cryptographic algorithms to detect changes in files and directories, some of which may be due to insertion of malicious programs that can cause systems and applications to crash or hang. Simply having software that can detect the presence of DoS tools does little good, however. It is necessary to not only systematically use this software at defined intervals, but also to ensure that the integrity of these tools is intact. Attackers love to modify these tools to make them unable to detect anything. Keeping these tools offline and loading them only long enough to run them is the wisest course of action.

Table 1 summarizes the types of preventative measures that can be used against DoS attacks and the relative effectiveness of each.

CONCLUSION

By now you should realize that DoS attacks constitute an extremely serious problem that has proven costly to many organizations. If anything, the problem is likely to become worse because systems, networks, applications, and the Internet itself are not really built to withstand these or any other type of attacks. With the increased threat of terrorist attacks also comes the likelihood that terrorist organizations will launch DoS attacks to accomplish their purposes. Yet the Internet community depends on continuous computing and network services. What can we do?

Unfortunately, there is no "silver bullet." Appropriate legislation would, however, be a good starting point. Government representatives need to take the threat of these attacks more seriously and must draft and pass legislation designed to more severely punish perpetrators of DoS and other kinds of cyberattacks. Considerably more funding for research on detection and prevention of DoS attacks would also constitute a big step forward. However, the ultimate responsibility is senior management's. Senior management must quit viewing attacks designed to compromise confidentiality and integrity as the "worse case" outcomes and must pay more attention to DoS attacks. Senior management must determine what the worst case DoS scenarios are and ensure that their organizations are prepared to deal with them.

The good news is that appropriate measures for preventing DoS attacks are not terribly different from those for dealing with other kinds of security-related attacks. Key measures include creating a sound policy and procedures, adopting the necessary technical countermeasures, being vigilant in detecting and analyzing anomalies, and responding promptly and effectively to incidents that occur.

GLOSSARY

Baseline controls An approach to information security in which an organization learns of the controls posture of other peer organizations and tries to deploy roughly the same controls posture.

Boomerang attack A type of DoS attack in which the attacker spoofs the IP address of the host that is the intended victim; the attacker sends packets that elicit some kind of reply, but if the attacker sends enough packets initially, a large volume of replies is returned to the victim, causing it to hang or crash.

Bubonic attack A type of attack in which a barrage of psuedo-random TCP packets with identical TCP sequence numbers, source and destination ports, and other packet header information is sent to a victim Windows 2000 system and certain versions of Linux, causing the system to crash.

Buffer overflow attack A type of attack in which a program receives too much input for the amount of memory allocated.

Christmas Tree attack An attack in which the attacker sends a packet with every option set for the TCP protocol.

Cold sites Operational sites to which computing operations are transferred after they are brought into operational status.

CPU Hog A Windows NT Trojan horse program that feeds its process threads at the expense of other, unprivileged threads, causing the system to crash several seconds after this program starts to run.

Denial of service (DoS) attacks Attacks are designed to disrupt, overwhelm, and/or damage computing resources and data.

Distributed denial of service (DDoS) attacks Attacks in which an attacker plants cooperative programs designed to unleash a massive DoS attack against a network.

Due care See baseline controls.

Failover systems and devices Redundant systems and devices that are immediately used if the primary system or device goes down.

Handler A type of program in a DDoS attack that controls zombies that have been planted in systems.

Hot sites Computing sites to which current computing operations can be rolled over in case of outage or disruption.

Jolt and Jolt2 attacks Attacks in which multiple identical fragments are sent to the victim system, causing massive CPU overutilization.

Land attack An attack in which an attacker sends one or more packets that have the same source and destination addresses, confusing the receiving system to the point that it crashes.

Master A type of program in a DDoS attack that controls handlers that have been planted in systems.

Packet fragmentation attacks Attacks in which packets are fragmented (broken into smaller pieces) in such a way that the receiving system processing them hangs or crashes.

Ping-of-death An attack in which a Windows 95/98/NT system is sent a series of fragmented IP packets, which when combined form a packet larger than the system can process, causing the system to freeze.

RAID (redundant array of independent disks) A hardware solution that allows uninterrupted access to data on disks if one disk fails.

Shaft A type of DDoS tool.

Shut-down attacks Attacks in which the purpose is to shut down but not delay systems or networks.

Slow-down attacks DoS attacks in which the purpose is to delay but not shut down processing activities of systems and/or networks.

Smurf attack An attack in which the attacker sends many spoofed ping packets to one or more systems within a network, causing the victim(s) to become overwhelmed with replies.

Spoofing Using an IP source address other than the real address of the machine launching this kind of attack.

SSPing An attack in which a Windows 95/98/NT system is sent a series of fragmented IP packets, which when combined form a packet larger than the system can process, causing the system to freeze.

Stacheldracht A type of DDoS tool.

TCP (Transmission Control Protocol) A protocol designed for reliable network communication between systems.

Teardrop attack A type of DoS attack in which the overlapping packet fragments cause the reassembled packet to have out-of-range values; if the receiving system is not programmed to check and drop fragments constructed in this manner, it will go into an abnormal operating condition and crash.

Tribe flood network (TFN) A type of DDoS tool.

Trinoo A type of DDoS tool.

TFTP (Trivial File Transfer Protocol) A protocol designed for efficient but not secure file transfer.

UDP (User Datagram Protocol) A protocol that is sessionless, connectionless, and unreliable.

Uninterruptable power supply (UPS) A hardware device attached to computing systems and network devices in a manner such that if an electrical failure occurs, the computer or device keeps running because of availability of UPS-supplied power.

Zombie A type of program in a DDoS attack that launches the attack after its handler instructs it to do so.

CROSS REFERENCES

See *Computer Security Incident Response Teams (CSIRTs); Computer Viruses and Worms; Disaster Recovery Planning; Encryption; Firewalls; Guidelines for a Comprehensive Security System; Internet Security Standards; Intrusion Detection Systems; Passwords; Physical Security; Secure Electronic Transmissions (SET); Secure Sockets Layer (SSL).*

REFERENCES

Gaudin, S. Net saboteur faces 41 months. *Network World*. Retrieved April 25, 2002, from http://www.nwfusion.com/news/2002/0304lloyd.html

Lemos, R. (2001). Study: Sites attacked 4,000 times a week. *CNETnews.com*. Retrieved April 25, 2003, from http://news.com.com/2100-1001-258093.html?legacy=cnet

National Academy of Sciences (2002). *Cybersecurity today and tomorrow: Pay now or pay later*. Retrieved from http://www7.nationalacademies.org/cstb/pub_cybersecurity.html

Schultz, E. E., & Shumway, R. M. (2001). *Responding to incidents: A strategic guide for handling system and network security breaches*. Indianapolis: New Riders.

Developing Nations

Nanette S. Levinson, *American University*

INTRODUCTION TO THE INTERNET AND DEVELOPING NATIONS

Today, 580.78 million people use the Internet. Only 5.12 million of these are in the Middle East; 6.31 million are in Africa; and 32.99 million are in Latin America. This compares with 167.86 million in the Asia/Pacific region, 182.67 million in the United States and Canada, and—the largest number—185.83 million in Europe (*How many online?*, 2002). Focusing in on specific countries within a region, one can start with South Africa where there are 238,462 Internet hosts or with Chad where there is one Internet host or Sudan where there are none. These data continue to indicate there is a global digital divide. Within developing nations themselves, there are large disparities between urban and rural areas and between elites and others in the country, making the gap between a developing country and other developed countries even more significant.

With regard to the language of the Internet, approximately 68.4% of the information on the Internet is in English. About 5.9% is in Japanese, with 5.8% in German, 3.9% in Chinese, 3% in French, and 2.4% in Spanish. Languages with between 1 and 2% each are Russian, Italian, Portuguese, and Korean. Other languages total 4.6% (Global Reach, 2002). Access issues clearly relate to the languages of content available on the Web. (It is interesting to note that the governments of China, Japan, and Korea recently agreed to a project promoting automatic translation of Web content into each other's languages to foster information sharing and Asian content on the Web for their citizens.)

Going beyond these numbers raises a set of very complex issues and challenges. To examine these issues calls for a systems-based perspective that recognizes and captures the possible interconnections among developed and developing nations as well as among governments, the private sector, international organizations, the growing nongovernmental organizational sector, and individual networks of experts and interested individuals such as those focusing on human rights or environmental issues. It also calls for new ways of highlighting such interconnections. No longer can we focus solely on a nation-state in discussing the Internet and development. The uneven yet increasing globalization and regionalization of economies, media, and knowledge provide an important backdrop.

Another area of change is what we mean by "developing" in the year 2003 and beyond. In the immediate post World War II era, "developing" was used to refer to the political development of a country. The model of development was the United States or other major Western democracies. This view of development placed the mass media in a key role in the development of democracy. Political development was viewed as separate from economic development. Little attention was paid to social development. Approaches to understanding economic development involved a stage theory of development whereby a foundation had to be laid for achievement of each succeeding step. Today, the words "development" and "developing" encompass a much broader definition, linking the individual to the nation-state and beyond. These words no longer convey only a stage approach. Rather, there are some instances of "leapfrogging" both social and economic development, especially using new information and communication technologies.

APPROACHES

There are four main analytic approaches to the Internet and developing nations. The first is rooted in the dependency paradigm, which has at its center a core–periphery relationship. Developed nations are at its core and developing nations are at its periphery; the core exploits and impacts the periphery (Van Rossem, 1996). The evolutionary approach argues that the environment selects out forms such as organizational forms that are to survive (Hannan & Freeman, 1988; Modelski, 1990). The institutional approach focuses on the role of institutions and institutional change in effecting development (Westney, 1993). It highlights three explanations for change: mimetic, normative, and coercive isomorphism (Powell & DiMaggio, 1991). Organizations or nation-states adopt a specific change because others are doing it (despite whether there are data regarding success of that change), because they ought to be doing it, or because some external power says they need to do it. Finally, an interorganizational approach captures the flow of resources (information, money, labor, raw materials, etc.) within groups of like and/or unlike organizations across complex organizational and cross-national boundaries.

This perspective allows for an understanding of a developing nation's potential to move beyond country-based competencies to the ability of its organizations to access and utilize a wide array of external resources and potential partners. Since the Internet itself is a network (a series of interconnected nodes), using an interorganizational perspective moves beyond a focus on core and periphery to incorporate a network of national and transnational factors and players. These include players such as the World Bank and the International Monetary Fund as well as multinational corporations, governmental organizations, nongovernmental organizations, and principled issues networks such as those focusing on human rights.

THE INTERNET AND DEVELOPING NATIONS: WHAT EXISTS

The Technology Achievement Index

Each of the above approaches can shape the views and policies of a developing nation regarding the Internet. No matter what the analytic approach, there are actions that developing nations can and do use to promote the Internet and foster innovations in support of development writ large. The 2001 United Nations Development Programme's *Human Development Report* presents for the first time a technology achievement index (TAI). This measure attempts to quantify how well a specific nation is doing with regard to four dimensions: creation of technology; diffusion of recent innovations; diffusion of old innovations; and human resource skills. Their measure of diffusion of recent innovations uses two indicators. First is the number of Internet hosts per capita (calculated from International Telecommunications Union data). Second is the amount of high- and medium-technology exports as a share of all exports (calculated from United Nations Statistical Division data). Telephones (both wired and cellular) and electricity are the indicators used to measure diffusion of old innovations. These measures are also useful in serving as foundations for understanding Internet use in developing nations.

The 2001 TAI has data for 72 countries. These data indicate that there are four main country groupings: first is leaders in technology achievements such as the United States and Finland. Second is potential leaders with high rankings on human resource skills and the wide diffusion of technologies but lower measures of innovation; such countries include Spain, Italy, and the Czech Republic. The third grouping is the dynamic adopters. These countries have high human resource skills, are home to high-tech industries, but have a history of slow or incomplete diffusion of old innovations. A number and, indeed, the majority of nations in this grouping are developing nations such as Brazil, China, India, Indonesia, South Africa, and Tunisia. The final grouping is called the marginalized set. Here, many parts of a country have not benefited from old technology being diffused nor is there evidence of the skills base for the diffusion of new technologies. Countries in this grouping include Nicaragua, Pakistan, and Senegal.

Public Policies

One key for developing nations with regard to the Internet is public policies. Such policies cover a range of areas and involve discussion with the private sector, nongovernmental organizations, and international organizations and institutions. Three themes emerge as central: transparency, stability, and adherence to the rule of law. A recent trend is the energetic involvement of the private sector in attempting to shape such policies. Associations of leaders of corporations and private sector associations especially in the technology sector have been formed to influence Internet-related policy making at both the nation-state and international organization levels. An example of such an initiative is the Global Business Dialogue formed in January of 1999 by chief executive officers of major companies involved in electronic commerce.

Currently, Internet costs are especially high in most developing nations. According to the 2001 *Human Development Report* (p. 81), the cost of being connected to the Internet in Nepal is 278% of the average monthly income there compared with 1.2% in the United States or 60% in Sri Lanka. Competition in Internet service providers (and privatization of telecommunications) is helpful in bringing down costs but needs to be combined with regulation, which helps to lessen the chance of private monopolies replacing the old public ones. Very recent data indicate a shift toward nation-states playing a larger role again in previously privatized or partially privatized telecommunications settings (Delaney & LaTour, 2002).

The promotion of research and development including fostering university–private sector efforts in research and development through, for example, matching funds or funds for science parks, also contributes to a more successful Internet utilization in developing nations. Additionally, policies that help small and medium-sized businesses and ensure entrepreneurship and venture capital make a difference. (Crafting a vibrant venture capital sector is not easy; it requires a relatively sophisticated financial system, ties to venture capital firms and entrepreneurs in other parts of the world, and strong educational systems—each of these elements presents big challenges, especially to the least developed nations.)

Much has been written about the role of education in fostering the Internet in developing nations, beginning with universal primary education and moving to solid secondary and tertiary education that includes strong verbal and mathematical training as well as teamwork and creativity enhancement. Recent research also highlights the need for excellent cross-cultural communication and trust-enhancing skills (Cogburn & Levinson, 2003). Finally, literacy is essential for effective Internet use in developing nations. This includes general literacy and computer literacy. Innovative approaches to computer literacy range from mobile kiosks for computer training in villages to cross-national alliances for Internet activities to provide vital knowledge to help small communities' economic endeavors. (Examples of alliances focused on developing nations include those which provide, via the Web, sophisticated weather information for fishing villages, Web sites promoting local crafts and obviating the need for third parties in the selling process, and distance education initiatives that recognize local cultures and often utilize local content on the Internet.)

Additional policies highlight training for current workers to support computer literacy and pave the way for successful innovation. Companies in many developing nations lack funds to support training efforts. Governmental policies providing even partial incentives and creative use of the Internet can help with training improvement. The UN *Human Development Report* recommends, with a view toward developing nation policy making, that a nation-state conduct inventories of skill levels, availabilities, and needs on a regular basis; not overlook helping small and medium-sized businesses, and provide information and incentives for such target businesses; and set up opportunities for recent secondary school graduates to get training (partially financed) in accredited private training centers for both urban and agricultural sectors.

These broad measures target the basic absorptive capacity or knowledge and skills base for a developing nation with regard to the Internet and related information and communication technology development. However, as the approaches summarized at the beginning of this chapter indicate, there are (depending on the specific analytic approach adopted) roles for developing nation governments in working with international institutions, regional organizations, and individual citizens. One example is a partnership between the International Telecommunications Union and Cisco Systems to provide Internet skills-related training in 50 new training centers in developing countries during 2003 including several centers in Africa, Saudi Arabia, and Tanzania focused on Internet skills development for women (ITU, 2002).

Secondly, there are policies that relate to the Internet and government functions within a developing nation. The term "e-government" comes into play here. From easy access to government information (including such things as databases on resources within a nation or even environment-related data) to services including electronic voter registration and voting, e- mail capabilities, and Web sites of elected officials and government offices, the Internet has great potential. Some governments have even developed Web portals, highlighting, in user-friendly ways, information resources and online services for their country as well as opportunities for foreign investment. (This potential is tempered by problems of access, literacy, and even possible corruption.)

Thirdly, there are policies that relate directly to Internet- and business-related research and development and cross-border Internet business. Countries with limited resources need to set priorities and goals within these far-ranging possibilities. One example is the government of Malaysia and its priority to develop a multimedia super corridor along with supporting policies and practices (Kushairi, 2002). Additionally, some governments are banding together at the regional level to pool scarce resources to address such priorities. In particular, there are regional efforts such as those in Asia to promote infrastructure for electronic commerce on a region-wide basis.

These efforts also relate to the promotion of interoperability and standards-setting that includes the participation of developing nations. Recently, the telecommunications ministers from China, Japan, and Korea discussed plans for a common wireless standard to promote economic development. Their actions demonstrate a regional approach to standards development in support of electronic commerce. (See ITU SPU, 2002a, for a report on the convergence of mobile telephony and the Internet.)

The policy approaches highlighted above focus mainly on strengthening economic development. There is another set of policy approaches evident in a minority of developing nations, which focuses on control of the Internet. A December 2002 study at the Harvard Law School examined China's possible blocking of Web sites over a six-month period (Zittrain & Edelman, 2002). It found that China had the tightest control of any other country during that period; it blocked access to 19,032 Web sites. Sites dealing with Taiwan or Tibet and portions of search portals such as Google were blocked at all or some portion of the time during the test period. The study also compared its findings on China with an earlier study it conducted on the blocking of Web sites in the Middle East. It reported that China exercised more broad but less in-depth control than Saudi Arabia.

Additionally, while China has shown great interest in promoting the growth of the Internet for economic purposes, it has pursued a multitiered strategy with regard to control. In July of 2002, a number of Web portals operating in China as well as Chinese businesses and research institutions signed a voluntary "Public Pledge on Self-Discipline for China Internet Industry." The pledge includes sections on preventing cybercrime and avoiding intellectual property infractions as well as on the blockage of any sites that would affect China's social stability or would bring harm to China (Bodeen, 2002).

The themes of intellectual property protection are particularly significant in the context of the Internet and developing countries. The World Trade Organization TRIPS agreement provides that the least developed countries have extra time until the year 2006 to implement systems of intellectual property rights using an agreed-upon set of minimum standards such as patent protection (United Nations Development Programme 2001, p. 103). There is still concern with the protection of the rights of communities and indigenous peoples. In addition, there is the need to balance protection with access to innovation.

Cybercrime is also an issue with regard to the Internet and developing countries. Some computer viruses have been tracked to individuals in developing nations such as the Philippines. Many developing nations have neither laws dealing with crime in cyberspace nor the technical skills necessary for investigative efforts. Since

11 September 2001, concerns about cyberterrorism have increased. There is significant discussion regarding the prevention of cyberterrorism and its possible forms and impacts as well as the need for international cooperation. There have been recent efforts to share expertise regarding cybercrime prevention and enforcement. In August 2002, the U.S. Department of Justice, the FBI, the U.S. Department of State, and AID held a training program in Moscow in conjunction with the Asia-Pacific Economic Cooperation Forum. This cooperative effort followed the signing of a counterterrorism declaration by the U.S. and Southeast Asian foreign ministers. The declaration particularly focused on strengthening and harmonizing laws related to cybercrime. The European Union has also been active in plans to combat cybercrime (*European Union Information Society,* 2002).

Finally, public policies need to focus on sustainability issues, ensuring that development with regard to the Internet becomes routinized and institutionalized, thus promoting long-term economic growth and positive social development. The systems approach (mentioned earlier), which calls for recognizing complex interconnections among state and nonstate actors, is particularly important in planning for sustainability. Policies that forge and reinforce embeddedness—links at different levels of groups of organizations—promote sustainability.

Barriers to Internet Development

There are at least 19 barriers that have been listed by Mansell and Wehn (1998) as hindering the development of the Internet and related capabilities in developing countries. There may be a very small domestic market and/or low levels of demand. Past government policy may not have included incentives for technology development. Perhaps companies have all been small and have not yet formed a network or alliances with companies from neighboring or from developed countries. There probably is no history of venture capital in the country. Nor is there easy access to credit for small and medium-sized businesses. Links between university, industry, and research institutes (if any) are weak or nonexistent. There is no relevant vocational training and an absence of flexibility and concomitant weak information management practices. Finally, there is little information and computer literacy among managers.

With regard to multinational and information and communication technology firms, there are few, if any, links. Labor costs within the country are high due to the lack of trained manpower. There is little expertise about a quality focus or about current software tools and development methods. The infrastructure is either weak or nonexistent or partially available in urban centers. Also weak is the knowledge base for marketing outside of a country. Finally, the government sector itself is a poor user of information and communication technologies in conducting its own business. Also it has few, if any, partnerships with the private sector.

Sometimes there is mistrust or a "not invented in this country" perspective, or there are problems in figuring out user needs, especially with regard to rural sectors or sectors with different cultures.

Understanding local culture is central. A major barrier is the absence of both local content and local language. What this often means is, then, an inability to promote social change and development. The software industry is located primarily in developed nations, although there are clusters within developing nations (such as Bangalore in India). Also, most software documentation is written in English—another barrier. A major problem is intellectual property rights violations and the absence of intellectual property laws in many developing nations. (See the earlier discussion of the World Trade Organization TRIPS Agreement.) Yet another hindering factor introduced earlier is the lack of ability to market outside of a developing country. While there are ways to deal with each possible barrier described above, many developing nations currently face these factors and are just beginning to grapple with them. The barriers appear to be greatest for the least developed nations.

STRATEGIES

Skills Inventories

Skills inventories need to highlight skills and knowledge resources both within and without developing countries' borders. Skills can be classified according to government workers, universities, research centers (if any), and the population at large. Particular attention needs to be paid to the education system from preschool through tertiary education and to training in the private sector (if any). Basic literacy is a foundation skill.

A second skill set is the ability to participate in networks and in knowledge sharing. This includes the ability to interface with computers, send electronic mail, and research on the Internet. Most importantly, it calls for a new way of thinking and interacting—an ability to think and work in network terms. Here, universities within a country or distance education partnerships can play a special role in developing a country's intellectual capital. A third skill is technical abilities that allow networks to function within a country, although this skill set is not required for the entire population. Vocational training programs can equip a subset of students with such expertise. Control and management skills are requisite for the survival of the Internet in developing nations. Such skills require budget and accounting expertise as well as managerial abilities and are required for effective government, nongovernmental organization, and private sector managers at all levels.

There is a fifth set of skills: leadership skills within government and within the populace. Leadership skills are vital to recognize where a nation is with regard to the Internet and information and communication technologies and to inspire a nation's advancement through, even partially, the Internet and related technologies. The ability to dialogue across generational borders and to communicate a vision to one's country and to leaders of all sectors as well as citizens is a vital and vibrant need. Also, most of all, leadership—a new kind of leadership—requires effective communication with leaders of other developing nations as well as with nongovernmental organization, international organization, and private sector leaders and other heads of government from around the world.

Developing Nations and Outside Factors

In addition to differences within developing nations and within regions, there are major differences with regard to Internet usage and policies as a result of factors from outside the borders of a given nation. A recent study of developing nations in five regions of the world finds that international forces are behind Internet usage within all countries in the study (Franda, 2002). At the same time, this study highlights the sovereignty of developing nations in their role as gatekeepers related to technology and innovation. The Internet has not undermined the power of state government. Rather, some developing nations evidence a pattern of limitation of Internet access and usage to specific elite segments of their population: the military, scientific, or other elites in a nation.

Evidence of 13 types of developing nation control from this study indicates specific limitations:

1. Laws that restrict or censor the Internet with regard to security, defense, pornography, criticism of leaders and their families, discussion of religion that might foster dissent, and related restrictions. (See the earlier-cited cases of China and Saudi Arabia.)
2. Laws that require registration with the government in order to access the Internet.
3. Penalties for putting items on the Web that violate a nation's laws.
4. Restriction of sites to which a citizen may gain access.
5. Shutting down of cybercafés where using the Internet may violate a nation's laws.
6. Restriction of uses of encryption technologies.
7. Restriction of licenses to universities and/or publishers and holding such institutions responsible for the behavior of those from within them for their usage of the Internet.
8. Requirements for Internet service providers to work with a government's intelligence and security agencies, thus giving such agencies access to e-mail messages and Internet content and then sharing with other governments' agencies.
9. Requirements that Internet service providers are responsible for Internet use and content that undermines national laws.
10. Allowing government intelligence agencies to review Internet e-mail and content.
11. Sending reminders or threats of possible disconnection of service in the event national laws are broken.
12. Making arrests related to violation of Internet-related laws.
13. Establishment of government monopolies over the Internet and related technologies to ensure complete control (Franda, 2002).

Additionally, there are reports of hackers at work breaking into targeted Web sites; of traffic analyses, tracking Internet messages and their flow; and of dissemination of incorrect or misleading information to cause confusion among dissident groups. In sum, privacy issues are central to a discussion of the Internet and developing nations. When attention to privacy issues is lacking, there is little opportunity for promoting the vision of open and electronic governance or—similar to the vision of the role of the mass media in the political development approach—the promotion of the Internet in building and strengthening democracy. Even attempts to work for Internet content in languages other than English can be viewed as opportunities for more, rather than less, local government control. The absence of policies regarding privacy protection and the Internet in many developing nations serves as a barrier to the development of an environment supportive of electronic commerce.

The Internet, Developing Nation Governments, and Diasporas

The Internet lends itself well to serving as a possible vehicle for linking more effectively homelands and diasporas as well as diasporas with no one single national home, especially in cases where there is an educated, dispersed population with Internet access. Diaspora-related Web sites can provide both socialization and identity-building for younger generations and also integration and identity reinforcement for older generations. The above-discussed Franda study provides information on how developing nation governments use the Internet successfully to link to diasporas (e.g., the worldwide community of individuals from India). Alternatively, diaspora communities themselves can design Web sites with no connection to a nation-state government.

There is much potential for governments to use diaspora Web sites to link homeland resources to external resources. Some countries have used Web sites to reverse a possible "brain drain," highlighting government policies (and related incentives) to lure former citizens back to their homeland. Others have used such Web sites for primarily political purposes—to maintain and enhance support for specific policies.

Developing Nations Policies: Overview

From a policy and nation-state perspective, there are many ways to consider what a nation-state ought to be doing. There are policies that relate to the Internet and business per se, including telecommunications, foreign direct investment, education, research and development, and tax policies. Some of these have been introduced earlier in the discussion of general public policies. One way to think about nation-states' policy roles is in the traditional terms of telecommunications policy, technology policy, industrial policy, and media policy. Each of these policy areas overlaps when it comes to the Internet and developing nations. However, the policy domain extends beyond individual countries' policy-making roles.

Building Infrastructure

A focus on building an appropriate infrastructure constitutes a basic strategy. A major challenge in many developing nations is not just the cost of access, as noted earlier, but just gaining access. There is a large gap in developing nations compared to developed nations even in terms of fixed-line telephones. In some countries, there are inordinate waits for a fixed-line telephone, even if the money is available to pay for the service. Electricity is often unreliable. One possible answer to the issue of the absence of fixed-line telephones and the unreliability of power sources is wireless. Recently, there has been a large growth in wireless, both in developed and developing nations. The 2002 International Telecommunications Union report on "Internet for a Mobile Generation" (ITU SPU, 2002b) highlights the convergence between mobile and the Internet and its potential for major economic growth. Yet a great deal of capital investment—especially for developing countries—is necessary. The report highlights Korea with its fast-growing mobile penetration rates as a possible model for a mobile information society in the future.

Another element is the development of policies that increase competition and bring down costs. Still another is the creation of innovative programs targeted at increasing access such as the program described next: The Grameen phone program that originated in Bangladesh provides an example of programs that build on local culture, provide opportunities for local entrepreneurs, and, most importantly, offer connections for local people at lower costs than fixed-line options. The Grameen Phone Program is modeled on the successful Grameen Bank Program. In fact, the Grameen Phone Program is spreading from Bangladesh to other developing countries. One example is the Grameen Phone Sewa Initiative, which is now about five years old. It focuses on villages in rural India. Escotel, a provider of cellular phones, makes one cellular phone available to each of a number of villages in rural India. The connections are free and airtime is subsidized. Operators recoup their investment in the phone in five to six months. However, most significantly, people in the village now do not have to walk days to make a call or to do business. Both Escotel and the Grameen initiative are doing well in India. These initiatives provide foundations for alternative access to the Internet in the future, given the convergence of Web and cellular technologies now being developed.

Information and Communication Technology Initiatives

A number of developing country governments have introduced information and communication technology initiatives as a central part of their technology policy programs. Malta is focusing on seven priorities: ensuring an information technology culture, promoting skill development related to information and communication technologies, promoting investment in information technology, helping organizations use information technology effectively, fostering an information technology industry within Malta that is export-oriented, using information technology effectively to deliver government services, and enhancing sector cohesion through information technology (Mansell & Wehn, 1998). Malaysia provides another example. Despite the Asian financial crisis, it continues its goal, as noted earlier, of a major global multimedia corridor and has thus far attracted over 500 companies from both inside and outside the country to its plan.

Another set of initiatives comes from the university sector. The Massachusetts Institute of Technology is now in its first year of implementation of its OpenCourseWare project (http://www.ocw.mit.edu). As of September 2002, it had placed 38 out of about several thousand courses on the Web. (This usually includes the syllabus, lecture notes, and any related materials.) MIT plans to complete the project in 2007. MIT reports that during the first month of operation, 315,000 people visited the site, with at least 30% from outside of the United States (Olsen, 2002). This project allows universities and individuals around the world access to MIT's detailed curriculum including computer science courses.

A second initiative involving the Massachusetts Institute of Technology along with Stanford and six other universities is the Open Knowledge Initiative (OKI). This project focuses on developing open (not proprietary) technical specifications for learning management systems (Web-based software that enables online class rosters, class outlines, assignments, discussions, and other classroom aids) (Olsen, 2002). The OKI project outcomes can assist developing nations in several ways: cutting development costs, and allowing organizations to take advantage of new applications with greater technical ease and increased system stability. There are several other similar attempts to share course syllabi and materials. What would be important as a next step is a mutual exchange, recognizing the knowledge base that developing country institutions may want themselves to contribute in a multidirectional learning experience. Another next step may be automatic or other translation of materials to match the language(s) of the country in which they are being used.

Guidelines for Initiatives

An important element for ensuring success of these strategies is the actual structure for dealing with such issues within the government. One of the major themes of the past decade has been both privatization of previous government monopolies of telecommunications and the separation of regulation and operation within government agencies. Setting up agencies that deal with regulation not operation of telecommunications, information, and media is a recent development in many developing nations.

The United Nations Commission on Science and Technology for Development has crafted guidelines for nations regarding information and communication technologies (Mansell & Wehn, 1998). First, there is a focus on information and communication technologies such as the Internet in social development. Great potential for social development exists through the effective use, for example, of the Internet as a tool for health education, prevention, and education. Second, there is a need to train and provide adequate human resources for the Internet and development-related initiatives. Third, there are new forms of organization and skills necessary for this information-intensive era. Each country needs, fourth, to have an effective information infrastructure with concomitant-trained human resources. Fifth, each country needs to focus on those most disadvantaged such as women and the poor in order to ensure appropriate training, access, and implementation. Sixth, countries need to consider partnerships with foreign companies and plans to stimulate direct foreign investment to meet market and growth needs. Seventh, there needs to be networks of expertise developed related to Internet and information and telecommunication technology development opportunities. Eighth, developing countries need to become key players at global tables determining standards, policy positions, and the like.

To carry out such guidelines, it is important that a country, as noted earlier, have an overall strategy for information and communication technologies. Such a strategy needs to recognize the regulatory and business environment worldwide. If there is to be a business environment conducive to markets, there must be stability, rule of law, and transparency with regard to policy. The costs of financing need to be considered as well as the need to serve the poor and rural areas. Participatory planning should be a key part of these considerations. Additionally, the strategy should address the development of human resources within the country. Issues of social development cannot be forgotten. At the same time, there needs to be a focus on ways to attract resources as well as on shaping a climate conducive to new business both indigenous and foreign.

The field of biotechnology provides an example for understanding the Internet and developing nations. Based on research in biotechnology, it is clear that the locus of innovation is no longer individual organizations. Rather, it is in networks of organizations and learning. Thus, developing nations need to think about not just a dyadic relationship with Internet and related technologies. They need to think about building relationships, informal or formal, with other developing nations, with other nations in a region, and with nongovernmental organizations. They also need to consider, as noted above, their relationships with their indigenous private sector and with foreign companies.

Examples of Initiatives

Developing nations' governments are working together, especially in response to the World Trade Organization. There are strong regional activities in Africa, Europe, and Asia, building upon active regional organizations such as ASEAN. There is also innovation. In September 2002, Egypt announced free Internet service, leading to growth in Internet use and access. Botswana is linking its legislators online and allowing citizens to follow the proceedings of their parliament. Linking to nongovernmental organizations, Bulgaria is working with municipalities in an anti-corruption initiative. Finally, in Sao Paulo, Brazil, the government is working with a nongovernmental organization to set up 100 telecenters to train people and small companies in computer usage. The South African government has established an information and communication technologies advisory board. This is an effective option for countries who wish to gain knowledge about the field yet retain their independence in policy making. An advisory board allows government officials to tap the knowledge resources of nongovernmental, university, think tank, and private sector organizations.

Even nongovernmental organizations are forming formal networks with one another and with governments of developing nations. One World is a partnership of nongovernmental organizations, formed to share knowledge. Recently, One World founded the Digital Opportunity Channel, working with another important category of organization, a foundation (in this case, the Benton Foundation) and with funds from two governments (United Kingdom and the Netherlands). The Digital Opportunity Channel is a Web site created to promote information flow and information sharing, about efforts to bridge the digital divide.

Yet another pattern (and developing nation governments often work with all of these aforementioned categories) is a developing nation government working with international organizations such as the United Nations Development Programme and international institutions such as the International Telecommunications Union. Starting in the 1990s, Estonia worked with the United Nations Development Programme to promote computer use by its citizens. Now, 100% of the new generation of Estonian students is computer literate.

Another major role is being played by groups of developed nation governments. The powerful Group of Eight has established a Digital Opportunity Task (DOT) Force and has also worked on an Africa Action Plan. The DOT Force is focusing on access, skill development, entrepreneurial enhancement, local content and applications, and health care. It recently announced an entrepreneurial network initiative. The DOT Force has established numerous partnerships including nongovernmental organizations and the private sector. For example, the DOT Force Entrepreneurial Network involves Accenture, Hewlett-Packard, Telesystem, and Open Economies (a policy center at the Harvard Law School).

Individual developed nations also have programs focused on providing assistance, especially to the least developed nations, with a focus on information and communications technologies and the Internet. For example, the U.S. Department of State through AID offers projects to assist governments with information technology-related development projects (U.S. AID Leland Initiatives, 2001). Additionally, the government of Japan provides a wide range of technology-related support from technical expertise sharing and technical training to professional exchanges (*New Breeze,* 2002).

Similar to nongovernmental organizations, the private sector is also learning to shape both informal and formal networks to promote learning, change, and their goals. As noted earlier in the chapter, heads of major multinationals from a range of countries have formed the Global Business Dialogue to promote electronic commerce. They are particularly focused on forming the policy framework for electronic commerce in emerging economies. Again, network organizational forms are key to the work of the Global Business Dialogue. Among its many partnerships, the Global Business Dialogue is working with e-ASEAN to focus on the necessary policy and legal infrastructure as well as a network infrastructure specifically for that region. It also signed an agreement to work on e-government and e-learning for the Asia-Pacific region. There are similar agreements with South Africa and with Egypt.

The work on e-government involves the following components, each of which has implications for any developing nation's use of the Internet for governance and government services: provision of online information (whether health or tax or governance-related); provision of online transactions and online services; design of an intranet within the government; design and assurance of security; design of electronic procurement; design of electronic learning; provision of adequate and secure technological infrastructure; provision of knowledge management; and design of electronic democracy.

The Global Business Dialogue also deals with cybersecurity issues and protection, essential to ensure the growth of electronic commerce in developing and developed nations alike. It is also thinking about the impact of electronic commerce, especially in terms of costs for small and medium-sized businesses. It works for increased and more effective cooperation between industry and government. Finally, it is promoting discussion and concern with cyber ethics.

There are numerous examples of nongovernmental organizations working not just with governments in developing nations, but also directly with individuals or informal groups. PEOPLink is a nongovernmental organization that used to sell items on the Web for artisans from a number of developing nations including Cameroon, Haiti, and Nepal. The idea was to cut out the middlemen. What is particularly fascinating is that recently PEOPLink announced that it is transforming its Web site so that, using the newest technology, artisans can sell directly online to people anywhere. This is an unusual e-commerce application that demonstrates the power of the Internet in transforming not just how Web sites operate but how artisans in developing nations may directly and effectively market and sell their wares.

Developing Nations and International Institutions and Organizations

There are a number of international institutions and organizations that are important in any discussion of the Internet and developing nations. Starting with two of the oldest (both a part of the United Nations Systems), the United Nations Development (UNDP) and the International Telecommunications Union (ITU) indicate ways in which organizations founded in eras without computer-related technologies have changed and adapted to this new telecommunications and Internet arena. Both UNDP and ITU have very active programs to work with developing nations, especially the least developed nations and with other partner organizations (and, especially increasingly over time, with the private sector). The UNDP has a networking and information technology observatory that deals with issues for developing nations (UNDP, 2002).

The ITU has numerous conferences and summits focused on Internet-related issues in developing nations. It highlights success stories related to the use of information and telecommunications technologies in developing nations. In 1995, the World Bank established a global grant program called InfoDev, which focuses on providing support of new information and telecommunication technology-related ideas to reduce poverty (*ICT Stories*, 2002). These organizations also provide knowledge resources related to the Internet and gender and the Internet and the environment.

An area where there needs to be more developing nations participation is in the area of global governance of the Internet. ICANN, the International Corporation for Assigned Names and Numbers, now manages the global system of domain names with the goal of maintaining the stability and also the security of the Internet. It has a Governmental Advisory Committee (GAC) with representatives of governments. Minutes of this Committee indicate that not enough developing nation governments send representatives to the ICANN meetings where decisions are made. (Naturally, there are costs involved to send representatives to such meetings.) Of particular importance for governments represented at ICANN is discussion of top-level country code domain names. Some developing nations such as Tuvalu have sold their country code domain name to a private sector organization (.tv in the case of Tuvalu) in order to raise funds for their country. Currently, ICANN is undergoing "evolution and reform" in terms of its structure and operations. The presence of a large number of developing nations in these deliberations would surely strengthen the outcomes.

CONCLUSION: FACILITATING FACTORS

In thinking about the Internet in developing nations, there are some important mechanisms at work that may be overlooked. The Internet itself as a technology or group of technologies that is, by design, a network facilitates the formation of additional networks. As noted earlier using the example of the biotechnology field, networks support the kind of learning that is important in this new information-intensive era. There is power through participation in networks to cut down on uncertainty and also to share what is tacit knowledge. Participating in alliances or partnerships or networks allows for complex learning (not just one-way learning or old-fashioned one-way diffusion of a model developed outside of a developing nation). Once networks are forged, their links are the conduits over which all sorts of resources including knowledge, influence, labor, materials, and money may be exchanged. In other words, networks of organizations and individuals promote interorganizational learning of many types and can build the absorptive capacity of organizations and even countries and regions.

One factor central to understanding the complex relationship between the Internet (or any technology) and any individual organization, country, or region, as well as sets of such, is culture. The cultures of participating organizations in a network and/or in a country or region shape and, in turn, are shaped by technologies as well as participants' cultures. Cultures influence how individuals and groups view concepts such as research, conduct negotiations, and make decisions. Many of the success stories highlighted by partnerships and networks focused on the Internet and developing nations come about because the participants recognize the roles of culture in their projects and in dealing with change. Change is a key element in development; Internet technology itself is central to the pace, depth, and nature of change.

Leaders of developing nations, international organizations, international institutions, nongovernmental organizations, private sector organizations, and related networks of organizations need to be able to design, manage, and respond to change. Most importantly, they need to be able to coach and to work well in networked situations. The Internet itself, by virtue of its information intensity and its technological makeup, is a vehicle of change—change that subtly but strongly impacts all elements and groups within its domain. Change can also be negative; today there are cases of cybercrime and cybercorruption in developed as well as developing nations. Thus, there is a need to ensure that government leaders as well as leaders of burgeoning private sector and nongovernmental sector organizations have the special skill set to deal with change in a networked world, and to propel development forward in an effective, ethical, and empowering way.

GLOSSARY

Absorptive capacity The degree of foundation present and necessary in an organization or country for using effectively new knowledge or new skills or new technologies.

Cybercafé A place where people come to use computers and access the Internet.

Diaspora Groups of people with similar ethnic and/or political identities who are dispersed, either voluntarily or not, around the world.

Diffusion of innovations The way in which an invention or a newly created process is disseminated and used.

Domain name The letters following the "dot" in an Internet address such as .gov, .edu, .com, or .org.

Electronic commerce An organization's or set of organizations' conduct of business functions via the Internet, which may be only business to business (known as B2B) or business to consumer, or both.

Electronic governance A country's use of information and communication technologies to conduct its business using the Internet, including access to government services and government information.

Globalization The increasing interconnection of elements such as the economy or knowledge or even nations as a result of the Internet and related technologies.

Infrastructure The physical (such as computers) and nonphysical (such as technically skilled human resources) elements necessary for the provision of basic information services in an organization or country and for subsequent growth.

Internet host A computer that allows many users to access network services (including the Internet) through it.

Interorganizational learning The process of identifying, acquiring, and using knowledge through being a part of a connected set of organizations such as those organizations participating in an alliance.

Isomorphism The processes of change which are either mimetic (through copying), normative (through doing what is right or what an organization ought to do), or coercive (through force of law or power).

Regionalization The process of increasing connections, both formal and informal, among a group of countries within a specific geographic region such as the Americas, Asia, Europe, or Africa.

CROSS REFERENCES

See *Digital Divide; E-government; Electronic Commerce and Electronic Business; Feasibility of Global E-business Projects; Global Diffusion of the Internet; Global Issues; Politics.*

REFERENCES

Bodeen, C. (2002, July 15). Internet portals in China sign pact to restrict access to information deemed subversive. Beijing: *Associated Press Worldstream.* Retreived April 4, 2003, from http://www.siliconvalley.com/mld/siliconvalley/3666643.htm?template=contentModules/printstory.jsp

Cogburn & Levinson (2003). US–Africa Virtual Collaboration in Globalization Studies: Success Factors for Complex, Cross-National Learning Teams. International Studies Perspectives, *4,* 34-51.

Delaney, K., & LaTour, A. (2002, September 9). Europe waffles on privatization as telecoms flail. *Wall Street Journal,* A1.

European Union Information Society (2002, November 20). Retrieved December 9, 2002, from http://europa.eu.int/information_society/index_en.htm

Franda, M. F. (2002). *Launching into cyberspace: Internet development and politics in five world regions.* Boulder, CO: Lynne Rienner.

Global Reach (2002). *Global Internet statistics: Sources and references.* Retrieved December 8, 2002, from http://global-reach.biz/globstats/refs.php3

Hannan, T., & Freeman, J. (1988, July). The ecology of organizational mortality: American labor unions, 1836–1985. *American Journal of Sociology,* 22–52.

How many online? (2002). Retrieved September 18, 2002, from http://www.nua.com/surveys/how_many_online/index.html

ICT Stories (2002). Retrieved September 18, 2002, from http://www.iicd.org/stories

International Telecommunication Union (ITU) (2002, December 3). *ITU and Cisco Systems to expand Internet training centre initiative to bridge gap in 'new economy' skills.* Retrieved December 8, 2002, from http://www.itu.int/newsroom/press_releases/2002/34.html

International Telecommunication Union (ITU) Strategy and Policy Unit (SPU) (2002a). *ITU Internet Reports 2002: Internet for a mobile generation.* Retrieved April 4, 2003, from http://www.itu.int/osg/spu/publications/sales/mobileinternet

International Telecommunication Union (ITU) Strategy and Policy Unit (SPU) (2002b, July–September).

Internet for a mobile generation. *ITU strategy and policy unit news update*. Retrieved April 4, 2003, from http://www.itu.int/osg/spu/spunews/2002/jul-sep/jul-septrends.html

Kushairi, A. (2002, October 28). Less developed nations catching up in Internet adoption. *Computimes Malaysia*.

Mansell, R., & Wehn, U. (Eds.). (1998). *Knowledge societies: Information technology for sustainable development*. New York: Oxford University Press.

Modelski, G. (1990, Winter). Is world politics evolutionary learning? *International Organization, 4,* 1–24.

New Breeze (2002, Winter). *14*(1).

Olsen, F. (2002, December 6). MIT'S Open Windows: Putting course materials online, the university faces high expectations. *Chronicle of Higher Education,* 31.

Powell, W., & Dimaggio, P. (1991). *The new institutionalism in organizational analysis*. Chicago: University of Chicago Press.

United Nations Development Programme (2001). *Human development report*. New York: Oxford University Press.

United Nations Development (UNDP) (2002). *Networking and information technology observatory*. Retrieved September 18, 2002, from http://www.sdnp.undp.org/observatory

U.S. AID Leland Initiatives: African Global Informational Infrastructure Project (2001). Retrieved September 19, 2002, from http://www.usaid.gov/leland

Van Rossem, R. (1996). The world system paradigm as general theory of development: A cross-national test. *American Sociological Review,* 508–527.

Westney, D. (1993). *Institutional theory and the multinational corporation. Organization theory and the multinational corporation*. New York: St. Martin's Press.

Zittrain, J., & Edelman B. (2002, December). *Empirical analysis of Internet filtering in China*. Retrieved April 4, 2003, from http://cyber.law.harvard.edu/filtering/china

FURTHER READING

Adamu, A. U. (2002, August/September). From minefield to mindset: Negotiating the information minefield in developing countries. *Bulletin of the American Society for Information Science and Technology, 28,* 25–30.

Edwards, S. (2002, May/June). Information technology and economic growth in developing countries. *Challenge,* 45, 18–43.

Hachigian, N. (2002, Summer). The Internet and power in one-party East-Asian states. *The Washington Quarterly.* 41.

Hudson, H. (1997). *Global connections: International telecommunications infrastructure and policy.* New York: Wiley.

Kamel, S., & Hussein, M. (2002). The emergence of e-commerce in a developing nation: Case of Egypt. *Benchmarking, 9,* 146–153.

Kowalczykowski, M. (2002, Summer). Disconnected continent. *Harvard International Review,* 40–44.

Kreb, B. (2002, August 14). U.S. aiding Asia-Pacific anticybercrime efforts. *Washington Post*.

Mahmoud, N. S. (2002). Determinants of Internet access demand in developing countries. *International Journal of Technology Policy and Management, 2*.

Schwartz, J. (2001, November 23). Cyberspace seen as potential battleground. *The New York Times,* B5.

Singh, J. P. (1999). *Leapfrogging development: The political economy of telecommunications restructuring*. Albany: SUNY Press.

Singh, J. P., & Rosenau, J. N. (2002). *Information technologies and global politics: The changing scope of power and governance*. Albany: SUNY Press.

Wilson, E. J. III. (2003). *The information revolution and developing countries*. Cambridge, MA: MIT Press.

DHTML (Dynamic HyperText Markup Language)

Craig D. Knuckles, *Lake Forest College*

INTRODUCTION

The *hypertext markup language* (HTML) is used to instruct a given piece of software how to "mark up" information for human viewers— the font sizes and faces, text colors, background colors, page layout, and so forth. A digital document composed of HTML is called an HTML document. An HTML document, which contains only plain text, is very different from its rendition, which is a visual display created by software according to the HTML markup instructions in the document. The rendition of an HTML document is called a Web page.

Many types of software can render HTML documents and create Web pages, but by far, Web browsers are the type of software most commonly used for that purpose. An HTML document stored on a local hard drive can be loaded into a Web browser and rendered as a Web page. Thus, the act of creating a Web page rendition from an HTML document has nothing to do with the Internet. The computer need not even be connected to the Internet. This is really no different from loading a word processing document into a piece of word processing software to see a rendition of that type of digital document.

The term hypertext refers to text with hyperlinks, or simply links. Text without links is simply text. HTML-rendering software, such as Web browsers, have the ability to retrieve and render new hypertext documents at the request of a user clicking a link. The rest of the name HTML designates it as a markup language. Markup instructions tell the rendering software about the document's logical structure and, to some extent, how its rendition should be formatted. HTML alone does not specify how humans can interact with the Web page. Hyperlinks provide for the only interactive capability of Web pages created using only HTML. That linking ability aside, a Web page rendition is inherently static. It can't interact with the user.

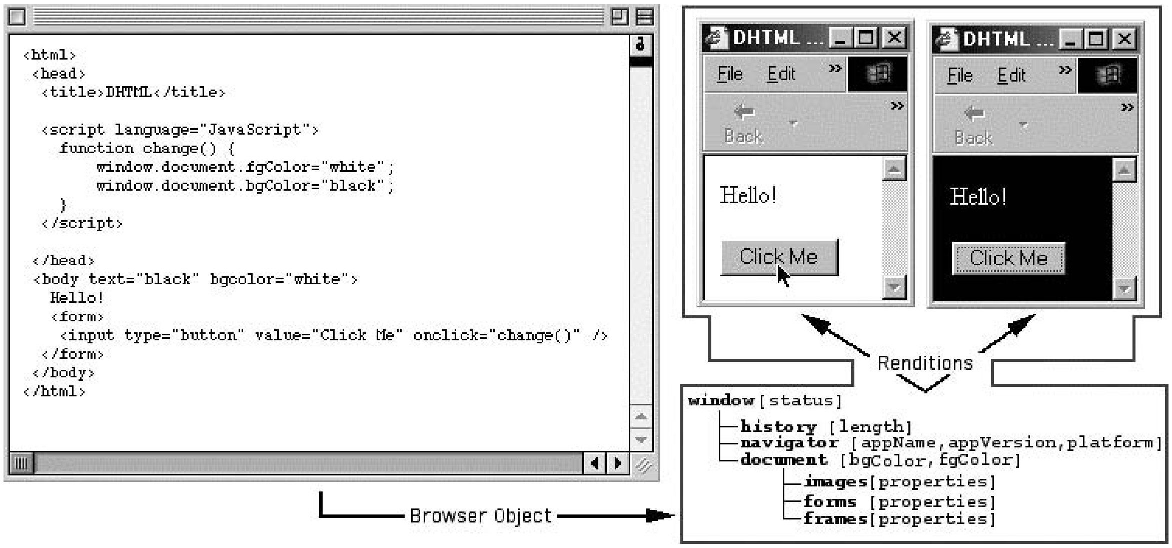
The reason HTML-created Web pages are inherently static stems both from the fact that HTML is only a markup language, albeit for hypertext, and from the fact that its transport protocol is *stateless*. The *hypertext transport protocol* (HTTP) is specified at the beginning of *a uniform resource locator* (URL), such as http://www.cubs.com, that requests an HTML document from over the Internet. The protocol is designated as stateless because it specifies that an HTML document, and objects such as graphics that are to be embedded in the Web page, be transferred to the Web browser and then that the connection to the remote Web server be terminated. HTTP does not provide for a continued transaction state between a Web browser and Web server.

That means that user interactivity can't be achieved in a rendered HTML-generated Web page through continuous support from the Web server. If fact, the legacy of an HTML document retrieved via HTTP from over the Internet is a Web page rendition that is no different from an HTML document loaded into the browser locally from the hard drive using no Internet connection whatsoever. Whether loaded locally or over the internet, a Web page rendition is static. A new connection to Web server is only invoked when the browser requests a new HTML document (or other resource) from the server. That just results in another static rendition in the Web browser.

Indeed, Tim Berners-Lee invented the World Wide Web around 1990 in an effort to mobilize information about current physics research on the Internet. The newly created HTML needed only to specify an electronic document's logical structure so that rendering software could understand how to display it. The newly created HTTP needed only to transport the documents to the rendering software over the Internet. Once retrieved, the HTML document provided for a static information display. Hyperlinks provided for a means to retrieve related information for static display. At the onset, Berners-Lee did not foresee that commercialization of the Web would lead to dynamic interactive effects in Web pages.

WHAT IS DHTML?

Dynamic HTML (DHTML) refers to a means for making a rendered Web page capable of customized interaction with the user without further Web server connections. A DHTML-enabled Web page has interactive capabilities that are handled entirely by the browser software on the user's computer. This is the abstract notion of DHTML. Typical DHTML interactive effects are triggered when a user passes the mouse over a region of the Web page and then something in the Web page changes. Examples of such changes include one graphic being replaced by another or a hidden layer in the Web page becoming visible. As the Web has evolved, Web page authors have sought to improve the functionality and aesthetics of a user's



```
<html>
 <head>
  <title>DHTML</title>

  <script language="JavaScript">
    function change() {
         window.document.fgColor="white";
         window.document.bgColor="black";
    }
  </script>

 </head>
 <body text="black" bgcolor="white">
   Hello!
   <form>
    <input type="button" value="Click Me" onclick="change()" />
  </form>
 </body>
</html>
```

Figure 1: A basic DHTML effect.

viewing experience. Indeed, a Web site that piques a user's interest is more likely to receive a return visit from that user. Through DHTML, the Web has achieved a level of interactivity sufficient to satiate many Web surfers' desires for instant interactive gratification, yet without taxing the Internet's infrastructure with superfluous, time-consuming network transactions.

Of all the Web-related technologies, DHTML is perhaps the most poorly named. A reference to "the dynamic hypertext markup language" leads one to think of a self-contained language, perhaps a big brother of HTML, used to enable dynamic interactive effects. But DHTML is not a self-contained language. A more accurate appellation for DHTML would be "dynamic Web pages using JavaScript." DHTML is actually a marketing term invented by the software vendors Netscape and Microsoft during the browser wars as they expanded on the dynamic capabilities of their browsers in efforts to gain market share.

The world has come to know DHTML as summarized by the following "equation."

$$\text{DHTML} = \text{HTML} + \text{CSS} + \text{DOM} + \text{JavaScript}$$

HTML is defined by the international standard W3C HTML 4.01 Recommendation. The *cascading style sheet* (CSS) language and the *document object model* (DOM) are defined by the international standards W3C CSS Level 2 Recommendation and W3C DOM CORE Recommendation, respectively. JavaScript, formally ECMAScript and defined by the international standard ECMAScript 262, is the scripting language used to provide programming support for Web browsers.

DHTML is a term that encompasses the ways the four distinct technologies on the right side of the above equation interact to create dynamic effects. There is no well-defined "language" specification for DHTML, nor even a definition of what DHTML should accomplish. Abstractly, DHTML enables interactivity in Web pages without the need for extra transactions with Web servers. Concretely, DHTML refers to what the imaginations of Web programmers have been able to accomplish by integrating HTML, CSS, DOM, and JavaScript.

The full strength of those four technologies need not be brought to bear to create DHTML effects. Very basic DHTML effects can be created using HTML, JavaScript, and a "subset" of the DOM, called the *browser object*. Such a very basic DHTML effect is illustrated in Figure 1. As this article progresses through the evolution of DHTML, its full strength will become apparent.

An HTML document is shown on the left of Figure 1. The body of the HTML document contains only one word and a clickable button. When the HTML document is loaded, the Web browser creates a browser object, which is the browser's in-memory representation of the Web page it creates from the HTML document and the window that contains it. That object is stored in the computer's RAM (random access memory) for quick access and holds the current state of the browser window and the Web page rendition. A partial depiction of the browser object is shown in the lower right of Figure 1.

The data about the current state of the window and Web page rendition is carried by the object in its *properties,* a few of which are shown in square brackets. Properties are basically ordinary programming variables, which are bound into the browser object's *hierarchical* structure. The hierarchical structure of the browser object is depicted in Figure 1. However, that depiction is a purely abstract visualization of the object. A browser does not display such a tree structure, but it makes the properties of the browser object available to the programmer through the hierarchy.

Properties of the browser object are accessed by using "dot notation" to progress down into the hierarchy. For example,

```
window.history.length
```

accesses the current window object maintained by the browser. A history object, which records information about the Web pages that have been displayed in the window, is associated with the window object. Its length property is a variable that contains the number of Web pages that have been loaded into the window since it was opened.

The variable

```
window.document.bgColor
```

is more pertinent to Figure 1. This refers to the document object (Web page) contained in the current window. The document object has several properties that contain data about the current Web page in the window. In particular, the bgColor property contains the page's current background color.

The left-most browser window in Figure 1 reflects the initial state of the document object as determined by the HTML instructions—black foreground (text) color and white background color. DHTML effects are created by changing the state of the browser object on a *user event,* in this case the user clicking the button. The onclick *event handler* defined for the "click me" button calls a JavaScript function named change(). That function is defined in a script included in the head section of the HTML document.

On the button-click event, the JavaScript function changes the state of the browser object by reversing the foreground and background colors of the document object. The effect on the Web page is immediate, and the user sees white text on a black background. The two browser windows in Figure 1 show the Web page before and after the button click. They do not represent two different browser windows. Rather, the DHTML effect triggered by the user event changes the state of the browser object and, hence, the Web page rendition. Moreover, the DHTML effect is instantaneous, with no Internet transaction involved.

The details of a specific example sometimes serve to overshadow the underlying concept, which can be quite simple when considered in more generality. Figure 2 shows the broader concept of DHTML implementation. This diagram doesn't reflect what CSS or the full DOM bring to DHTML. Those additions add nothing to the general concept, just more flexibility and many more details regarding implementation.

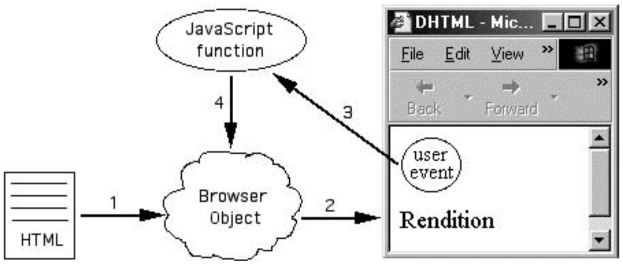


Figure 2: The general concept of DHTML. 1. The browser parses the HTML document and creates an in-memory (RAM) browser object, which stores the properties of the Web page as specified by the HTML document. 2. The Web page rendition reflects the current state of the Browser object. 3. A user event causes a JavaScript function to be executed by the browser. The JavaScript function is in a script contained in the HTML document. 4. The JavaScript function changes the state of the browser object. This causes an instantaneous change in the state of the Web page rendition.

Steps 2 through 4 as shown in Figure 2 need not be a one-time sequence of events. In fact, a sequence of user events can trigger a sequence of DHTML effects: event calls function -> function updates state of browser object -> rendition state changes -> event calls function -> repeat process. As the result of repetitive user events, DHTML can actually create animations in a Web page.

The types of user events that trigger DHTML effects are varied. For example, the act of the user loading a Web page into a browser (the onload event) can trigger DHTML effects from the onset. But most often, DHTML effects are triggered by the user sometime after the page is fully rendered. Some common events, which are user actions for which the browser is basically sitting there "listening," are listed below.

Onclick—The user clicks the mouse on a certain region of the Web page.

Onmouseover—The user moves the mouse over a certain region of the Web page.

Onmouseout—The user moves the mouse back out of a certain region of the Web page.

Onmousemove—The user merely moves the mouse.

The DHTML effects initiated by such events vary widely.

When defining a technology, it is often instructive also to define what it is not. Java Applets and Flash animations are the most common technologies that can be misconstrued as being DHTML. Java Applets and Flash animations are basically small applications (hence the term Applet) that can be embedded in a Web page. In both cases, the Web browser requires a helper application to execute the embedded technology. Java Applets require a Java virtual machine (JVM), and Flash animations require a Flash plugin for the browser. Both of these technologies increase the runtime overhead of the browser and are not favored for simple visual DHTML-style effects. The increased overhead for the browser, and the fact that a given user might have to install the helper application just to see the effect, are the main reasons why DHTML is favored for relatively simple interactive enhancements. But when interactive needs become elaborate, such as for small video games embedded in Web pages, Flash and Java are certainly worth the extra overhead they incur.

Many books have been written on how HTML, CSS, DOM, and JavaScript work together to create DHTML. A book by Knuckles (2001) provides such coverage from the ground up and assumes no prior programming experience. A book by Teague (2001) provides more elaborate coverage of specific effects that can be created using DHTML.

THE EVOLUTION OF DHTML

At the dawn of Web time (circa 1990), Web pages contained only text and hyperlinks. The initial goal of HTML, and its transport protocol, HTTP, was to mobilize information on the Internet. There was no initial intent for fancy visual effects. The goal was a Web of information, interconnected via hyperlinks. That goal has been realized far beyond the original expectations. The Web now mobilizes billions of HTML documents.

Nonetheless, the first 5 or so years of the Web's existence featured only Web pages with hyperlinks. The Web did not explode until Marc Andreessen developed an application, called Mosaic, that enabled Web pages to contain graphic images: hence its name. This was the first piece of software that had functionality resembling that of what we today call Web browsers. Andreessen went on to form Mosaic Communications Corporation, which was later renamed Netscape Communications. The release of Netscape Navigator version 2 for all major computing platforms (Microsoft's Windows, Macintosh, and Unix) caused the Web to ignite. The evolution of DHTML begins there.

Netscape 2 (March 1996)

Netscape Navigator 2 (NN2) introduced JavaScript 1.0, which provided the first scripting support behind Web pages. Prior to this, the only interactivity provided by Web pages was clicking hyperlinks. Now, programmers could write scripts to respond to user events and provide simple interactivity.

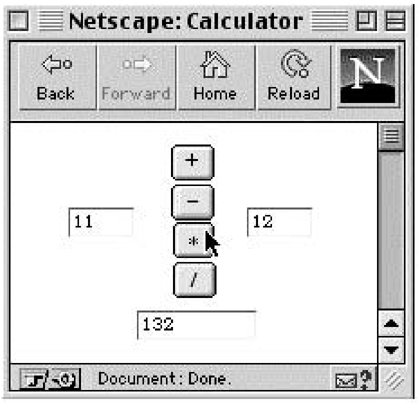
Client-side processing of user data was one important feature enabled by JavaScript 1.0. The browser object supplied by NN2 was sufficient to allow JavaScript 1.0 to manipulate data entered into HTML forms. The simple calculator utility rendered in NN2 shown in Figure 3 demonstrates that. Only the relevant parts of the HTML document and browser object are shown. Remember that this is purely an abstract depiction of the hierarchical object maintained in RAM by the browser. From this, it is apparent how the JavaScript references to the objects are formed using the dot notation.

When the HTML document is loaded into the browser, the browser object is initialized to reflect the current state of the document. The browser object hierarchy reflects the containment relationships. The window contains the document. The document contains the form, whose name is calcForm. The form contains the text fields, whose names are num1, num2, and result. The value properties of the text fields contain the data entered into the fields. In this case, all three text fields in the Web page are initially empty when the page first loads.

In Figure 3, the user has entered two numbers and clicked the multiplication button. The JavaScript function, which would be included in a script in the HTML document, first retrieves the two numbers from the browser object. It then calculates the result and assigns that result to the third text field. When the user clicks the button, the update to the browser object is instantaneous, and the answer immediately appears in the third text field. Again, no extra transactions with the server are involved.

Data processing on the client has become a staple of Web programming. Real Web applications use a mixture of client-side processing and server-side processing to process user information, such as personal information, credit card numbers, and airline flight choices. It is interesting to note that this is not considered DHTML by many, although it meets the abstract definition of DHTML given above. Colloquially, the term DHTML is generally used to refer to fancy visual effects rather than data processing. Nonetheless, the addition of HTML forms and JavaScript in NN2 introduced the world to the Web as a computing platform rather than just a means of displaying information.

The browser object of NN2 did allow for some effects that are sometimes placed under the umbrella of DHTML, although the term was yet to be coined. Those effects were mostly limited to changing the state of the browser window. Examples include changing the tool bar, status bar, or size of the browser window. Figure 4 shows a simple



part of Browser Object

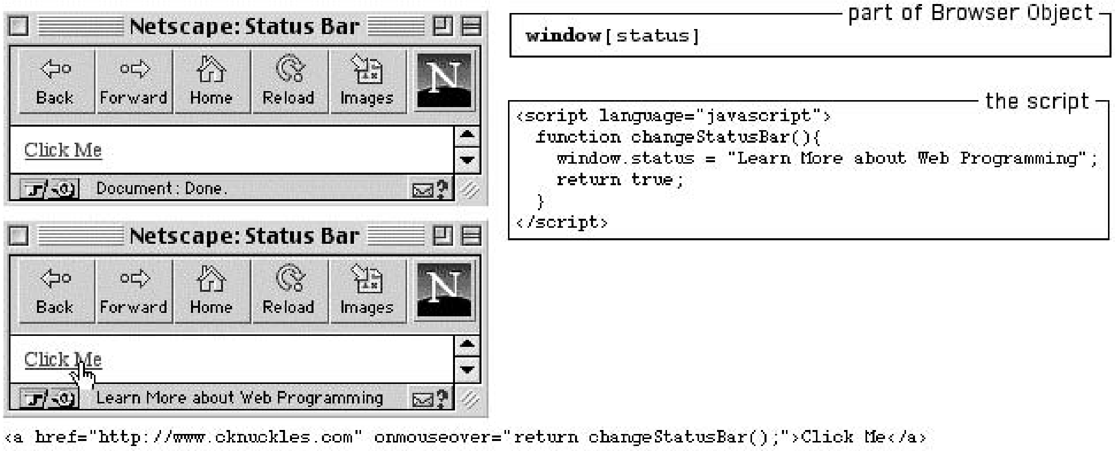
```
window
 └─document
     └─calcForm
         ├─num1[value]
         ├─num2[value]
         └─result[value]
```

the script

```
<script language="javascript">
  function calculate(operation){
    var num1=parseFloat(window.document.calcForm.num1.value);
    var num2=parseFloat(window.document.calcForm.num2.value);

    var result = eval(num1+operation+num2);

    window.document.calcForm.result.value=result;
  }
</script>
```

```
<form name="calcForm">
  <input type="text" name="num1" size="5" />
  <input type="button" value="+" onclick="calculate('+')" />
  <input type="button" value="-" onclick="calculate('-')" />
  <input type="button" value="*" onclick="calculate('*')" />
  <input type="button" value="/" onclick="calculate('/')" />
  <input type="text" name="num2" value="" size="5" />
  <input type="text" name="result" value="" size="10" />
</form>
```

Figure 3: A calculator in a Web page in Netscape Navigator 2.



Document: Done.

Learn More about Web Programming

```
window[status]
```

```
<script language="javascript">
  function changeStatusBar(){
    window.status = "Learn More about Web Programming";
    return true;
  }
</script>
```

```
<a href="http://www.cknuckles.com" onmouseover="return changeStatusBar();">Click Me</a>
```

Figure 4: A simple interactive effect in Netscape Navigator 2.

"DHTML" effect in NN2 that places a description about the nature of a hyperlink in the status bar when the user passes the mouse over the link. Only the relevant parts of the HTML document are shown.

The browser window in the upper left of the diagram shows the link, which gives no indication of to what Web page it points. The second browser window shows the result of the onmouseover event. The status bar is at the very bottom of the browser window. The JavaScript function called by the onmouseover event simply changes the status property of the window in the browser object to create this effect.

Although this effect is more like "dynamic browser window" than DHTML, these types of interactive visual effects are what DHTML is inherently about. It was not until around 2 years after NN2 was released that the term DHTML was coined in earnest. But simple browser window effects would come to be part of what DHTML encompasses. In the mean time, Netscape Navigator 3 and Internet Explorer 3 would add yet more to the foundation that would become DHTML.

Netscape 3 and Explorer 3 (Late 1996)

The Web was moving very quickly. In the same year they released NN2, Netscape released version 3 of their browser (NN3). Also, Microsoft entered the browser market with their release of the Internet Explorer 3 (IE3) software. For all practical purposes, this was the first commercial version of their browser, the 3rd-version designation being intended to indicate their software was on equal footing with NN3.

With NN3, Netscape released JavaScript 1.1, and with IE3, Microsoft released JScript. JScript is a scripting language reasonably similar to JavaScript. Web developers tended to stick to the common subset of JavaScript and JScript, which was adequate for most scripting purposes. NN3 dominated the browser market at this time, and JScript would never really gain its own identity.

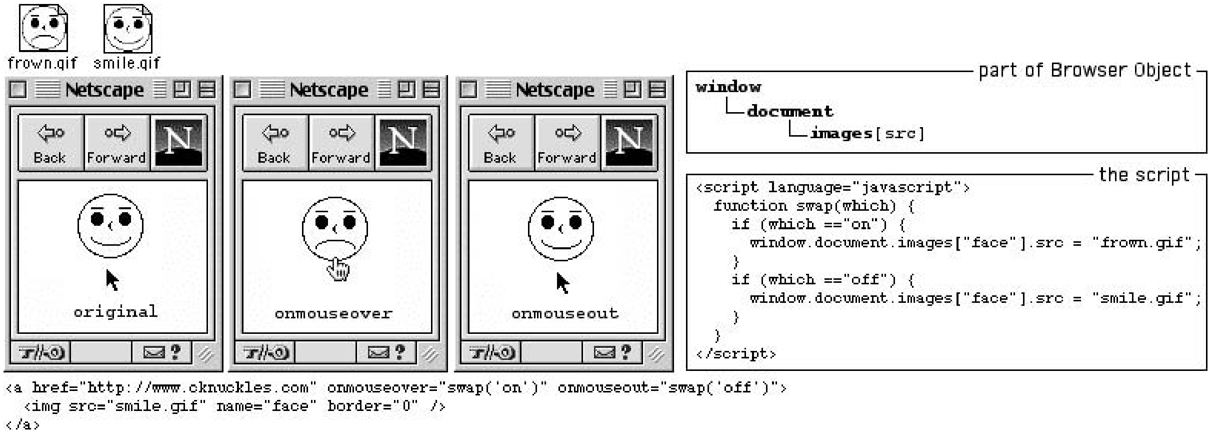
The biggest difference between the two Web browsers was that they featured substantially different browser objects. Again, Web programmers tended to stick to a common subset of the two browser objects to achieve cross-browser compatibility. The commonality between the browser object of NN3 and that of IE3 became what the world would come to know as *the* browser object. Figure 1 depicts part of the browser object, which is actually much more extensive.

The representation of HTML forms in the browser object, as shown in Figure 3, became the mainstay for data collection and manipulation in Web browsers. Although that capability was present in NN2, NN3 and IE3 featured the robust form manipulation capabilities that are still an integral part of e-commerce data-collection activities today. The browser object did, and still does, provide for full manipulation of text fields, checkboxes, radio buttons, pop-up menus, and all other form elements to which we have grown accustomed.

But the browser object only allows for a few DHTML effects. The "windowing" effects, most notably the pop-up windows that typically contain advertisements, were inherited by both browsers from NN2. The most notable new addition that NN3 and IE3 added to the visual effects arsenal was rollover graphics. The rollover graphic effect is created when the mouse rolls over (onmouseover) a graphic embedded in a Web page, causing that graphic to be replaced with a different one. When the mouse rolls back off the graphic (onmouseout), the original graphic is substituted back into place. This effect is most commonly used for graphics that are active as hyperlinks in Web pages. The fact that the mouse causes the image to change indicates to the user that the image is, in fact, active as a link.

Figure 5 shows how a rollover graphic is accomplished. As shown at the top of the figure, two distinct graphic image files are used, smile.gif and frown.gif. When the HTML file is first loaded into the browser, the original image on display is the smiley face. When the user passes the mouse over the image, the onmouseover event handler calls the swap() JavaScript function to "turn on" the rollover. When the mouse passes off the graphic, the onmouseout event handler turns the rollover back off by returning the original smiling image. As the two images are swapped in and out of the browser object, it simply appears to the user that the face changes.

Both swaps are accomplished by changing the source of the image in the browser object. Figure 5 only shows



```
window
  └─document
      └─images[src]
```

```
<script language="javascript">
  function swap(which) {
    if (which =="on") {
      window.document.images["face"].src = "frown.gif";
    }
    if (which =="off") {
      window.document.images["face"].src = "smile.gif";
    }
  }
</script>
```

```
<a href="http://www.cknuckles.com" onmouseover="swap('on')" onmouseout="swap('off')">
  <img src="smile.gif" name="face" border="0" />
</a>
```

Figure 5: An image rollover in Netscape Navigator 3.

the relevant part of the browser object. As specified by the HTML code, the name of the image in the browser object is face. When the JavaScript function changes the source of the image, the change is instantaneous from the user's perspective. But for the change to be instantaneous, the graphic that is to replace the one originally on display needs to be downloaded from the Web server at the time the Web page is retrieved. If the replacement graphic is not "preloaded" into the browser's memory, then an extra transaction with the Web server at the time the rollover is triggered is necessary to retrieve it. That defeats the purpose of browser-handled interactive effects. The JavaScript code necessary to preload the replacement image is not shown in Figure 5.

The most prevalent use of rollovers on the Web has been, and continues to be, for navigation tabs and bars. In these situations, there are typically several images that are swapped in and out. It is not uncommon for rollover displays to use 10 or 20 different graphics. Figure 6 shows two typical rollover displays. The horizontal navigation bar (nav bar) on the left features several menu tabs. The upper picture shows the menu bar in its initial state, indicating that the user is viewing the home page. The lower picture shows the menu bar when the mouse passes onto the preface tab. When the mouse is moved back off the preface tab to a location not on the nav bar, the display reverts to its initial state. Or, if the mouse is passed onto a different tab, that tab becomes highlighted. When the user actually clicks on a tab, a new page is loaded containing the same nav bar, but the default highlighted tab then corresponds to that page.

Such nav bars are often comprised of several images, one for each tab, which fit together seamlessly. The rollover display then requires an active, and a nonactive version of each image. Thus, the nav bar on the left of Figure 6 would require 12 different graphics, and each rollover action would replace two tabs—make one inactive and one active. Alternately, the whole nav bar could be one graphic. That requires one version of the nav bar for each active tab. That would require only six different graphics in Figure 6. A rollover action would swap the entire graphic in or out depending on which tab the mouse passed over. The user would not notice the difference between a display that uses six graphics and one that uses 12.

The vertical nav bar shown on the right on the right of Figure 6 could come from an online tutorial, for example. Such nav bars often just highlight the particular link over which the mouse passes, but this one swaps in arrow graphics pointing to the mouseover link. The central graphic is divided into an image map of different regions, one region for each lesson link. There is a stack of a blank white graphic on either side of the central graphic. This rollover effect does not replace the central graphic when the mouse passes over it, but it replaces the white graphic on either side of the mouseover region with an arrow graphic. There are only four graphics used to create this rollover display: the central graphic containing the lessons, the blank white image that is replicated in stacks on either side of the central image, and the two arrow graphics that are swapped in for the appropriate white "dummy" images.

When the mouse is moved up and down over the different lessons in the central graphic, the arrows move with the mouse, creating a simple animation-type effect. It is up to the imaginations of Web designers how rollovers can be employed to create novel DHTML effects. At the

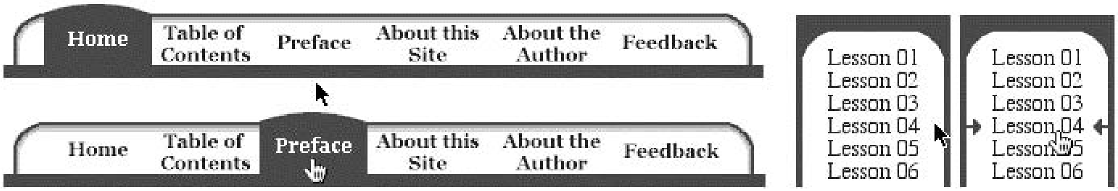


Figure 6: Typical uses of rollover graphics.

time of NN3 and IE3, effects such as those shown in Figure 6 were considered very fancy and sophisticated in Web pages. Multiple rollovers were at the pinnacle of Web page design.

This was also the era in which "browser sniffing" was born. Web programmers began to insert into Web pages JavaScript code that would "sniff out" the type of browser that was viewing the page. The properties of the navigator object in the browser object, shown in Figure 1, contain information about the browser's vendor, its version, and even the platform (Mac, Windows, Unix/Linux) on which it is running. (Ironically, that object is still named navigator, even in current Explorer browsers.) Once the JavaScript code determines the particular browser with which it is dealing, it can get the most out of the particular browser object.

At this time in Web history, the cross-browser multiple rollover is the effect of choice and is, in fact, the seed that sprouted into DHTML. It is not entirely clear exactly when the term materialized, but it was right around the time that rollover graphics became widely popular. It is clear, however, that the term proliferated as marketing jargon used by the browser vendors Netscape and Microsoft to tout the dynamic interactive capabilities of their newly released version 4 browsers.

Netscape 4 and Explorer 4 (1997)

Microsoft and Netscape were desperately locked into the "browser wars" at this time. Netscape was the spark that had ignited the Web, but Microsoft ruled the majority of desktop computing with its Windows operating system. It was that edge that eventually won the war for Explorer. But at this point in time, the battles still raged on.

In efforts to capitalize on the new idea of DHTML, Netscape Navigator 4 (NN4) introduced a new HTML layer element, with which multiple layers of content could be superimposed in one browser window. In contrast, Internet Explorer 4 (IE4) pushed the use of CSS and the *z*-index to create multiple layers. Regarding a flat browser window as a two-dimensional *x–y* coordinate plane, the *z*-index adds a third dimension. Netscape's HTML layer element also adds a third dimension, but not using CSS.

Both layering techniques turned the once "flat" Web page into three-dimensional *x-y-z* coordinate space. It's not that layers could appear to "pop out" at the user, but one layer could be moved independently of other layers by a JavaScript function as the result of a user event. Such effects are the cornerstone of "modern" DHTML. Prior to layers, a Web page layout was like a picture frame with multiple holes into which pictures could be inserted. The holes were in fixed locations in the page, and one could only swap images of the proper size in and out of a hole. This analogy explains rollover graphics rather well.

A layered Web page is more like a stack of loose pictures on table. One can shuffle the pictures around, leaving some partially covered, completely hiding some, and revealing others that were not previously visible. A picture can move in its own layer, independently of pictures in other layers. Hiding, revealing, and moving layers is the basis for much of "modern" DHTML.

But due to the radically different DHTML implementations by the two browsers, "modern" DHTML was largely a lost cause in this era. Industrious Web designers constructed rather elaborate JavaScripts to sniff out browsers in order to move layers in both implementations. But the majority of Web sites stuck to rollover graphics. Indeed, the Web had moved quickly in terms of browser versions, and a significant portion of the Web still used version 3 browsers. Most any Web site wanted their pages to work properly in everyone's browser. Many sites even avoided rollover graphics in this era because they wanted to sell products to those still using NN2.

Modern Browsers

It turns out that CSS and the *z*-index used for layers would be the future of DHTML. CSS, which adds to HTML the ability to exactly control the markup of Web pages, enables advanced page layout and many new and exciting DHTML effects. CSS is fully supported (for the most part) by modern Web browsers. That includes IE5+, NN6+ (a NN5 version was never released), Opera 4+ (a browser gaining popularity in Europe), and Mozilla 1+ (from the open source Mozilla project). These browsers are compatible (again, for the most part) due to the emergence of accepted international standards for Web technologies. These standards have enabled a whole new generation of DHTML effects. Modern DHTML is discussed below, following a brief overview of the international standards that enable the modern DHTL effects across different computer platforms and browser types.

THE STANDARDIZATION OF DHTML

Although it's not well-defined as a distinct technology, DHTML has become predictable because the technologies of which it is comprised have been internationally standardized. DHTML's aggregate technologies became uniform international standards beginning around 1998. The standards continue to evolve, but the following core standards on which DHTML is currently based are stable and widely supported by modern Web browsers (NN6+, IE5+, and Opera 4+).

HTML 4.01: This World Wide Web Consortium (W3C) standard has been stable since 1999. One goal of HTML 4.01 was to keep the language small, relegating many Web page formatting tasks to CSS (see W3C HTML 4.01 Recommendation, 2002; Graham, 1998, 2000; Knuckles, 2001).

CSS Level 2: This W3C standard has been stable since its revision in 1999. The main goals of CSS are to provide for elaborate formatting potential for Web pages, and to keep rules for how a Web page should look separate from the data that comprises the Web page. Most "modern" DHTML effects are created by changing style properties after the Web page is fully loaded (see W3C CSS Level 2 Recommendation, 2002; Sklar, 2001.

ECMA Script: This standard, developed by the European Computer Manufacturing Association (ECMA) and approved by the ISO/IEC (International Organization for Standardization/International Electrotechnical

Commission), has been stable since 1999. It is usually called JavaScript and is based mostly upon JavaScript 1.1. It does contain some features from JScript (see ECMAScript 262 Standard, 2002; (Gosselin, 2000; Knuckles, 2001).

DOM: DOM is a standardized API (application programming interface) that exposes an electronic document's logical structure to programming languages. This is a W3C standard, most of it stable since 1998, the rest since 2001. Different levels of the DOM specification are relevant to DHTML (see W3C HTML DOM CORE Recommendation, 2002; W3C HTML DOM Recommendation, 2002).

DOM Level 0 (DOM0): DOM0 is implicitly defined as the common subset of the browser objects of NN3 and IE3. DOM0 is the browser object that has been used for all DHTML examples so far in this chapter. This is the API used ubiquitously for manipulating HTML forms and creating older DHTML effects, such as window manipulation and rollover graphics.

DOM Levels 1 and 2 (just DOM): Exposes the complete logical structure of a document to a programming language. DOM is quite general and is used to expose XML (extensible markup language) documents to various programming languages (JavaScript, Java, Perl, C++, etc.). DOM is used for an API in many applications, not just Web browsers.

A specialized subset, called *HTML DOM,* is geared for HTML, JavaScript, and Web browsers. HTML DOM is backward compatible with DOM0, meaning that the browser object is a "subset" of the HTML DOM. HTML DOM adds the ability to manipulate CSS properties in a Web page using JavaScript.

MODERN DHTML

DOM0, the browser object, is sufficient for such things as manipulating windows, manipulating HTML forms, and creating rollover images. Expanding on DOM0, HTML DOM exposes CSS style properties to JavaScript and features "tree traversal" methods used to access HTML elements. Traversing an object tree can be done using full reference through the hierarchy. For example, the long DOM0 references used in Figures 3 and 5 are window.document.images['imageName'].src and window.document.calcForm.result.value.

In DOMO, objects such as forms and images are found at predictable locations in the object tree, so accessing them using full object reference is not a problem. However, the HTML span and div elements, which are generic containers that allow CSS styles to be applied to specific parts of Web pages, can be nested unpredictably in HTML documents. That means that it is more difficult to find such objects in the DOM tree using full reference from the tree's root.

The HTML DOM allows an HTML element, such as span or div, to be assigned unique ID (identification). Then an element can be directly "plucked out" of the DOM object tree, regardless of its location within the hierarchy, using it's unique ID. For example, suppose an HTML document contains a block division that is used to apply a CSS style class named myStyles.

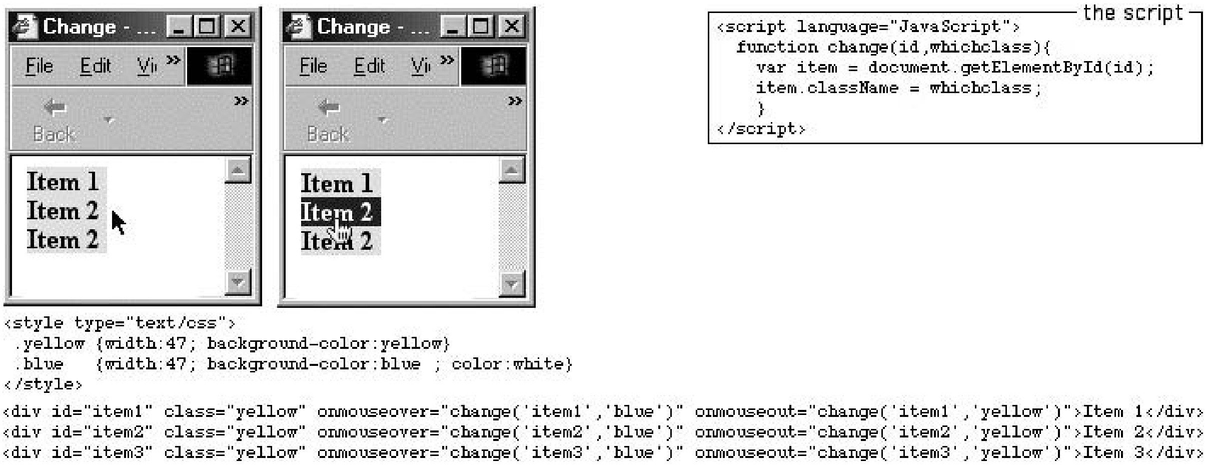
```
<div class="myStyles" id="block1">content
  of block </div>
```

Regardless of where the div element appears in the HTML document, hence regardless of how "deep" it may be in the DOM hierarchy, HTML DOM allows it to be accessed directly.

```
document.getElementById("block1")
```

Once accessed in a script, the CSS style properties of the block can be changed to create DHTML effects. But note that using the HTML DOM to access styles in this way is only viable in the most current browsers (i.e., version 6+ of NN and IE).

Three simple DHTML effects can be used to explain many of the most useful effects. Those simple effects are changing the appearance of an object, moving an object on the screen, and changing its visibility status. A simple appearance change is demonstrated in Figure 7. The menu is created from three div blocks, each assigned the yellow style class. That class features black text on a yellow

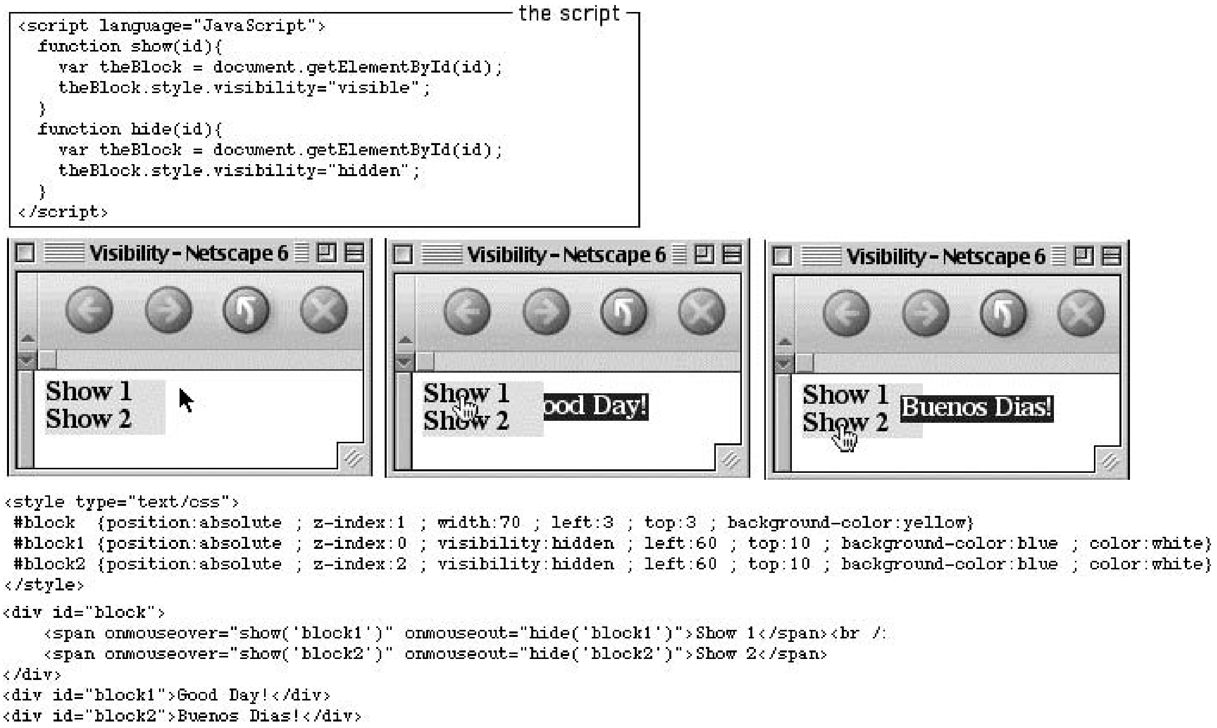


```
<script language="JavaScript">
  function change(id,whichclass){
    var item = document.getElementById(id);
    item.className = whichclass;
    }
</script>
```

```
<style type="text/css">
 .yellow {width:47; background-color:yellow}
 .blue   {width:47; background-color:blue ; color:white}
</style>
```

```
<div id="item1" class="yellow" onmouseover="change('item1','blue')" onmouseout="change('item1','yellow')">Item 1</div>
<div id="item2" class="yellow" onmouseover="change('item2','blue')" onmouseout="change('item2','yellow')">Item 2</div>
<div id="item3" class="yellow" onmouseover="change('item3','blue')" onmouseout="change('item3','yellow')">Item 3</div>
```

Figure 7: Changing CSS classes.



```
<script language="JavaScript">
  function show(id){
    var theBlock = document.getElementById(id);
    theBlock.style.visibility="visible";
  }
  function hide(id){
    var theBlock = document.getElementById(id);
    theBlock.style.visibility="hidden";
  }
</script>
```

```
<style type="text/css">
 #block  {position:absolute ; z-index:1 ; width:70 ; left:3 ; top:3 ; background-color:yellow}
 #block1 {position:absolute ; z-index:0 ; visibility:hidden ; left:60 ; top:10 ; background-color:blue ; color:white}
 #block2 {position:absolute ; z-index:2 ; visibility:hidden ; left:60 ; top:10 ; background-color:blue ; color:white}
</style>
<div id="block">
    <span onmouseover="show('block1')" onmouseout="hide('block1')">Show 1</span><br /:
    <span onmouseover="show('block2')" onmouseout="hide('block2')">Show 2</span>
</div>
<div id="block1">Good Day!</div>
<div id="block2">Buenos Dias!</div>
```

Figure 8: Changing CSS visibility properties.

background. When the mouse passes onto one of the blocks, onmousover calls a JavaScript function that changes the style class of that particular block to the blue class. The blue class features white text on a blue background. When the mouse passes back off of the block, onmouseout again calls the function, which changes the block's style class back to its original yellow class. Figure 7 shows both the onmouseover and onmouseout states of the menu. In appearance, this effect is similar to that created by rollover graphics.

Figure 8 demonstrates the dynamic effect of changing the visibility status of each of two div blocks. The example also demonstrates layering with the CSS *z*-index. When this page first loads, only the main menu appears in the browser window. The styles for the other two blocks, block1 and block2, are set so that those blocks are not visible when the page first loads. When the mouse passes over the "show 1" span of the main menu, onmouseover calls a JavaScript function that causes block1 to become visible. Because block1 has a *z*-index of 0, it is partially obscured by the menu block, whose *z*-index is 1. The blocks are positioned on the page in such a way that the layering is evident due to the overlap of the blocks.

When the mouse passes back off of "show 1," onmouseout calls a function causing block1 to become hidden again. Similarly, passing the mouse onto "show 2" causes block2 to become visible. But block2 partially obscures the menu block because its *z*-index is 2.

One common DHTML effect enabled by layering and visibility control is pop-up text. Figure 9 shows a hyperlink with extra text support that pops up when the mouse passes over the link. When the mouse passes back off the link, the pop-up text disappears. Moreover, as shown in the middle and on the right (Figure 9), the location of the popup text block depends on the location at which the mouse passes onto the link. That technique involves "capturing" the exact position of the mouse on the screen at the time of the onmouseover event. The JavaScript code for that is much too involved to show here.

There are at least two ways the pop-up box shown in Figure 9 can be accomplished. First, the block can be toggled back and forth between a *z*-index that is below the main content layer of the Web page and one that places it above the main content layer. Second, the most common method, is simply to change the visibility status of the pop-up block depending on whether the mouse is over the link or not. A given Web page may have many blocks not initially visible, either because of layering or because of visibility status, that pop up on user events.

Going beyond text that merely supports navigation, entire navigation menus can be caused to pop up on user events. Those menus can themselves contain navigation links. Figure 10 demonstrates that effect. When the user passes the mouse over (or clicks on) one of the menu tabs at the top of the Web page, a menu appears. The user can

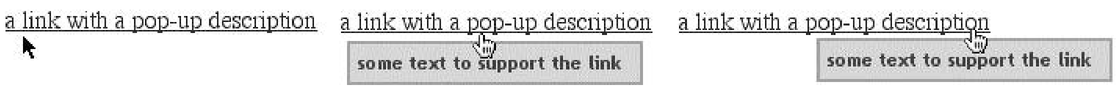

a link with a pop-up description
a link with a pop-up description
some text to support the link
a link with a pop-up description
some text to support the link

Figure 9: A hyperlink with pop-up text.

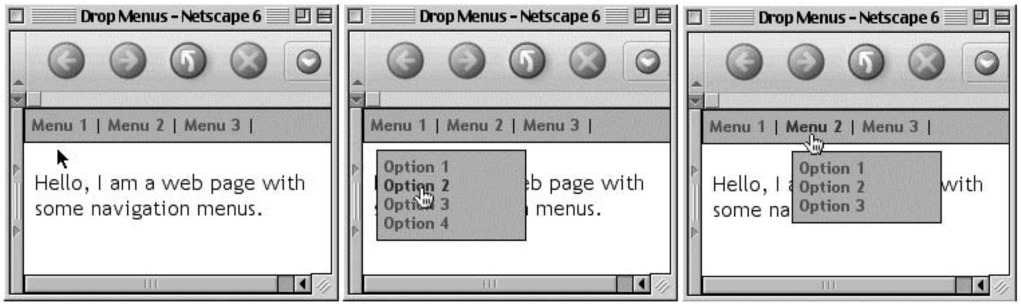

Drop Menus - Netscape 6
Menu 1 | Menu 2 | Menu 3 |
Hello, I am a web page with some navigation menus.
Option 1
Option 2
Option 3
Option 4

Figure 10: Pop-up navigation menus.

then select a link from the menu to follow. If they chose not to follow a link in the menu, passing the mouse off of the pop-up menu causes it to disappear. In functionality, this DHTML effect is similar to the drop-down menus provided by most software in a Windows environment.

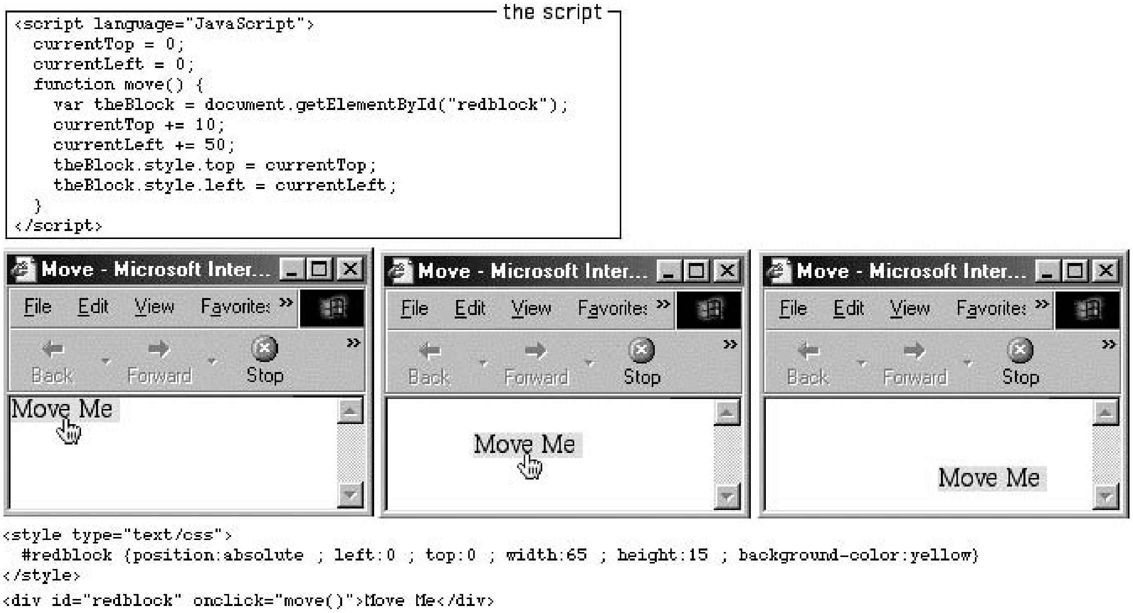
Figure 11 demonstrates the movement of blocks using DHTML. When the page first loads, there is one yellow div block positioned on the page in the upper left corner of the browser window. Clicking on the block invokes a JavaScript function that adds 10 to the *y*-coordinate of the block and 50 to the *x*-coordinate. When those new coordinates are assigned to the appropriate style properties of the block, its position on the screen moves instantaneously. Figure 11 shows the original block position and then its position after two movements caused by mouse clicks.

All the code aside, the concept demonstrated in Figure 11 is quite simple. A user event calls a JavaScript function that moves a block by a certain increment. A sequence of events, in this case mouse clicks, causes a sequence of movements. It is possible for one user event to trigger an entire animation. This is accomplished when the function that causes the incremental movement is called repetitively in an automated fashion. The function below, given in pseudo code, demonstrates how that works.

```
function moveBlock() {
  move the block by some increment
  if ( the block should move again ) {
     setTimeout("moveBlock()", 100)
  }
}
```

Again, the concept is fairly straightforward. Some event calls the moveBlock() function, which moves an object by some small increment. If the block is supposed to keep moving (i.e., the condition that stops the movement is not met), then the moveBlock() function is automatically called after a timeout. Here the timeout is 100 ms (0.1 s). The “speed” of the animation is determined by the timeout delay. When the movement increment is small and the timeout is relatively short, the repetitive function calls cause repetitive incremental movements of the object, thereby creating an animation that appears continuous on the screen.

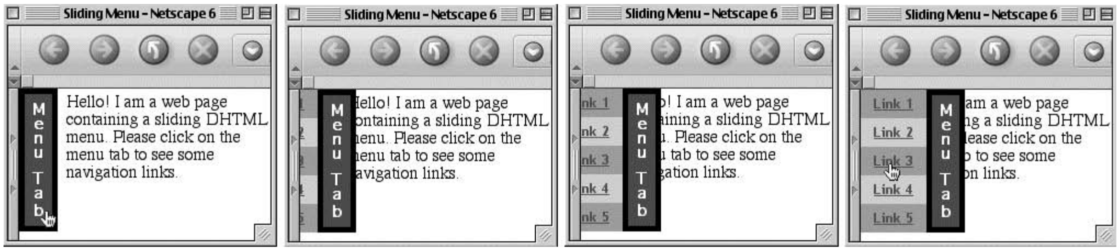
Figure 12 shows a practical application of DHTML animation. When the page first loads, all the user sees is a Web



```
<script language="JavaScript">
  currentTop = 0;
  currentLeft = 0;
  function move() {
    var theBlock = document.getElementById("redblock");
    currentTop += 10;
    currentLeft += 50;
    theBlock.style.top = currentTop;
    theBlock.style.left = currentLeft;
  }
</script>
```

```
<style type="text/css">
  #redblock {position:absolute ; left:0 ; top:0 ; width:65 ; height:15 ; background-color:yellow}
</style>
<div id="redblock" onclick="move()">Move Me</div>
```

Figure 11: Moving objects using DHTML.



Hello! I am a web page containing a sliding DHTML menu. Please click on the menu tab to see some navigation links.

Figure 12: A sliding menu.

page with a menu tab protruding from the upper left corner. A mouse click on the menu tab triggers a JavaScript function that repeatedly calls itself using a timeout. The menu then moves relatively quickly onto the screen until the condition of the far left portion of the menu reaching the edge of the window stops the animation. Reclicking on the menu tab causes the menu to recede from the screen in a similar fashion. In a regular-size Web page, that tab might be only be an inch or two (3 to 5 cm) tall and would not dominate the Web page, as it does in this illustration.

Many DHTML effects are not so practical and are intended for show. One such example utilizes layers to create a rather novel effect. The page shown in Figure 13 features black text on a black background. Hence, no text is visible without help. The page background has a *z*-index of 0, and the text in the page has a *z*-index of 2. The help comes from a white graphic, whose *z*-index is 1, and which follows the mouse. As the mouse moves closer to the text in the page, the white graphic moves under the black text (*z*-index 2), but over the black background (*z*-index 0). Thus, the text becomes visible to the extent that the white graphic has passed under it.

Many DHTML effects that feature things that follow the mouse around the screen are much more elaborate than that shown in Figure 13. However, the basic principle is the same. On each onmousemove event, the on-screen coordinates of the move are captured from the DOM using a JavaScript function. Then, the function updates the coordinates of the object(s) following the mouse to reflect that movement. In practice, mouse-following effects are rarely used because often such movement distracts and annoys users to a greater degree than it amuses them or provides added functionality to the Web site.

Many DHTML scripts can be downloaded for free on such sites as Dynamic Drive and webbedENVIRONMENTS. Some are relatively simple and, after a careful reading of the instructions provided, can easily be adapted to fit specialized needs. Others are extremely elaborate. Those can be used as is but require a knowledgeable programmer to modify. Some are practical, such as navigation menus, but many are just novelties, as shown in Figure 13. In the end, the possible DHTML effects in Web pages are limited only by the imaginations of Web programmers.

THE FUTURE OF DHTML

As of 2002, when this article was written, the "modern" DHTML effects described above were just starting to find widespread use on major commercial sites. Major sites can't afford to lose business from their pages not working properly in all browsers. Thus, content development usually lags at least one browser version behind the most current ones. The technologies might evolve fairly quickly, but widespread implementation is contingent on the diminishing market share of older Web browsers.

One CSS feature that bodes well for the future prospects of DHTML is the pseudo-style class, the most notable being the hover pseudoclass for hyperlinks. The hover pseudoclass features a built-in onmouseover event handler that changes the style properties of the hyperlink to those specified by the hover pseudoclass when the mouse passes over the link. This DHTML effect uses no JavaScript whatsoever. Figure 14 demonstrates two hover links. This effect is viable in version 5+ browsers.

The hyperlink on the left in the figure is initially not underlined. When the mouse passes over the link, its CSS properties are automatically changed to a hover pseudoclass, which specifies underlined text. In contrast, the hyperlink on the right is initially underlined. When the mouse passes over that link, it's hover pseudoclass causes it to become rendered as white text on a black background.

It is possible that future revisions of the CSS standard will include more pseudoclasses that feature built-in event handling and interactivity. It is also possible that Web browser vendors will extend the hover pseudoclass

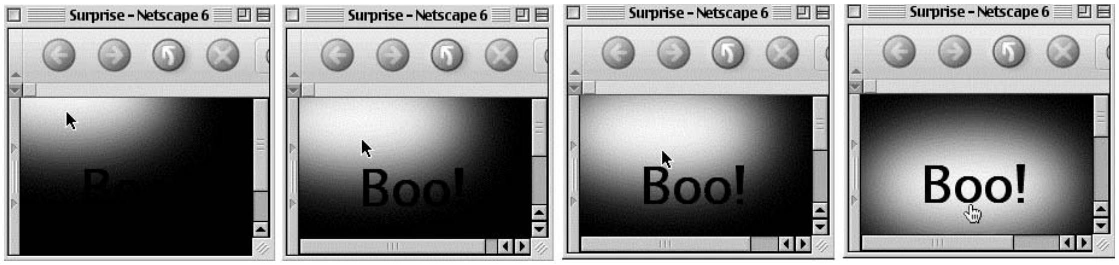


Figure 13: Following the mouse.

Click Me
Click Me
Click Me
Click Me

Figure 14: CSS hover links.

to HTML elements other than the anchor element, which creates hyperlinks. For example, if an automated hover pseudoclass could be assigned to a generic div container, the DHTML effect shown in Figure 7 could be accomplished without using JavaScript. When software vendors implement practical innovations, the World Wide Web Consortium (W3C) is often not far behind in creating an international standard, provided standardization is warranted.

The *scalable vector graphics* (SVG) W3C recommendation (W3C Scalable Vector Graphics Recommendation, 2001) also bodes well for the future of dynamic effects in Web browsers. Scalable vector graphics are like those that can be created using the "draw" utility of word processing software. Also, they have become a staple layout feature for such software as Microsoft's PowerPoint. These easily created and manipulated graphics soon will be viable in Web pages. The SGV recommendation features DOM access and a JavaScript binding that will allow DHTML style effects using vector graphics. Of course, it will take several years before major Web sites employ such technology. DHTML will likely take on a whole new look by the time version 6 browsers are considered outdated.

CONCLUSION

Some definitions of DHTML regard it as including only the modern, CSS-enabled effects demonstrated above. Indeed, the term DHTML was invented around the time HTML DOM first enabled those more elaborate effects. However, when defining DHTML according to what it accomplishes and how it is implemented, as opposed to defining it in marketing terms used to tout new features added by browser vendors, it takes on more breadth and scope.

DHTML effects are implemented when JavaScript code, usually called in a function as the result of a user event, changes the state of the Web browser's in-memory object representation of the Web page. Changing the state of the browser's object, whether DOM0 or HTML DOM, changes the Web page rendition. Under this general definition, common effects enabled by DHTML include browser window manipulation, image rollovers, hidden layers that pop up, and objects that move in Web pages. Technically, the manipulation of data in HTML forms fulfills the general definition of DHTML, but that capability is usually classified as client-side data processing in Web browsers.

DHTML should not be confused with embedded technologies such as Java Applets and Flash. Those technologies require helper applications, usually called browser plugins, to execute in Web browsers. These technologies are much more powerful than DHTML, but they use more resources on the client computer and may even require intermittent or continuous interaction with the Web server. DHTML is, and will continue to be, favored for adding fast and efficient interactive capabilities to Web pages.

For Web developers who do not wish to write customized JavaScript code or develop highly customized DHTML effects, numerous Web sites offer free DHTML scripts and tutorials on how to use them. Moreover, many software packages automatically create DHTML effects in a visual environment. Macromedia's Dreamweaver and Fireworks are the two most notable applications with that capability. Adobe's GoLive is also worth mentioning. But such software often generates cumbersome and inefficient JavaScript code. Worse yet, proprietary software often generates code that is not browser neutral. Nonetheless, DHTML-creating software allows inexperienced Web developers to quickly create interactive DHTML-enabled Web pages.

GLOSSARY

Browser object A Web browser's in-memory (RAM) representation of a Web page.

Cascading style sheets (CSS) Provides precise formatting and layout instructions to augment HTML.

Document object model level 0 (DOM0) The browser objects used by Internet Explorer 3 and Netscape Navigator 3.

Dynamic HTML A collection of technologies and techniques that allow Web pages to be interactive without further transactions with a Web server.

ECMAScript The official standard, called JavaScript in common parlance.

Event When a user moves the mouse or clicks on a region of a Web page.

Event handler A trigger in place to deal with user events. Event handlers are usually used to call JavaScript functions.

HTML DOM An extension of the browser object.

Hypertext markup language (HTML) Instructs software how to render a Web page.

Hypertext transport protocol (HTTP) Stateless protocol that coordinates transport of HTML documents on the Internet.

JavaScript A scripting language used to provide programming support for Web browsers.

Pseudoclass A CSS style class that has built-in behavior. The hover pseudoclass creates DHTML link rollovers.

Rendition Interpretation and display of an HTML document by software. The rendition of an HTML document is called a Web page.

Rollover A DHTML effect triggered when the mouse passes over a region of a Web page. Rollover graphics refers to rollovers causing graphic images to be replaced in Web pages.

Scalable vector graphics (SVG) Will play a big part in the next generation of Web page design and DHTML.

***z*-index** The layer-creating CSS property. Usually used to layer HTML div containers.

CROSS REFERENCES

See *Cascading Style Sheets (CSS); HTML/XHTML (HyperText Markup Language/Extensible HyperText Markup Language); JavaScript; TCP/IP Suite.*

REFERENCES

Dynamic drive. Retrieved April 19, 2002, from http://www.dynamicdrive.com

ECMAScript 262 standard. Retrieved April 19, 2002, from http://www.ecma.ch/ecma1/STAND/ECMA-262.HTM

Gosselin, D. (2000). *Comprehensive JavaScript*. Boston: Course Technology.

Graham, I. S. (1998). *HTML 4.0 sourcebook*. New York: John Wiley & Sons.

Graham, I. S. (2000). *XHTML 1.0 language and design sourcebook*. New York: John Wiley & Sons.

Knuckles, C. (2001). *Introduction to interactive programming on the Internet*. New York: John Wiley & Sons.

Sklar, J. (2001). *Designing Web pages with cascading style sheets*. Boston: Course Technology.

Teague, J. C. (2001). *DHTML and CSS for the world wide Web*. Berkeley: Peachpit Press.

W3C CSS level 2 recommendation. Retrieved April 19, 2002, from http://www.w3.org/TR/REC-CSS2/

W3C DOM CORE recommendation. Retrieved April 19, 2002, from http://www.w3.org/TR/DOM-Level-2-Core

W3C HTML 4.01 recommendation. Retrieved April 19, 2002, from http://www.w3.org/TR/html4

W3C HTML DOM recommendation. Retrieved April 19, 2002, from http://www.w3.org/TR/DOM-Level-2-HTML

W3C scalable vector graphics recommendation. Retrieved April 19, 2002, from http://www.w3.org/TR/SVG

webbedENVIRONMENTS. Retrieved April 19, 2002, from http://www.webbedenvironments.com/dhtml

Digital Communication

Robert W. Heath Jr., *The University of Texas at Austin*
Atul A. Salvekar, *Intel, Corp.*

INTRODUCTION

Digital communication is the process of conveying digital information from a transmitter to a receiver across an analog channel. The origin of the binary data is known as a *source;* the destination of the binary data is known as a *sink*. While binary data may be derived from an analog source such as music or a digital source such as a Web page, the means by which the binary data were created has little influence on the operation of the digital communication system. Digital communication could also be defined for nonbinary sources, but for current transmission systems this is not standard.

The principles of digital communication have been recognized and rediscovered many times during the past few thousand years. Early forms of digital communication used technology such as smoke signals, torch signals, signal flares, or drums. Most of these systems were visual, meaning that the message was conveyed through sight by signaling according to some prearranged code. One of the more successful signaling systems is the heliograph, discovered in ancient times and still in use today, which uses reflections of the sun from a small mirror to convey digital signals. Digital communication using electrical signals is more recent and dates back to the invention of the telegraph by Samuel Morse in the 1830s. The telegraph used Morse code, which is essentially a mapping from letters to quaternary sequences (long pulses, short pulses, letter spaces, and word spaces), to convey digital information over long distances via cable. Marconi patented a wireless telegraph in 1896—this is the origin of wireless digital communication. The facsimile (fax) machine is a sometimes surprising example of early digital communication. First patented in 1843 by Alexander Bain, the fax machine both then and now scans documents line by line and digitally encodes and conveys the presence or absence of ink.

Despite its early beginning, digital communication fell out of prominence with the invention of the telephone in 1875. The telephone is an example of an analog communication system that conveys a continuous source (in this case speech) instead of digital information. With the availability of computers beginning in the 1960s, and the resulting need to exchange digital data over great distances, interest in digital communication and digital communication systems reemerged.

Digital communication systems offer a number of advantages over comparable analog systems. A basic advantage is that they are fundamentally suitable for transmitting digital data. Digital communication systems, however, offer other advantages including higher quality compared with analog systems, increased security, better robustness to noise, reductions in power, and easy integration of different types of sources, e.g., voice, text, and video. Because the majority of the components of a digital communication system are implemented digitally using digital signal processing, digital communication devices take advantage of the reductions in cost and size enjoyed by digital signal processing technology. In fact, the majority of the public switched telephone network, except the connection from the local exchange to the home, is digital.

All aspects of the Internet are enabled by digital communication technology. The backbone of the Internet is digital communication over optical fibers. All last-mile access technologies, despite the different transmission media, are fundamentally digital including voice-band modems, cable modems, and digital subscriber lines. Local area networks use different digital communication technologies such as IEEE 802.3 (Ethernet) for wired access or IEEE 802.11 for wireless access. Metropolitan area networks, such as IEEE 802.16, use digital communication to cover larger geographic areas. Digital communication allows remote access to the Internet via cellular systems through CDPD in first generation systems and GPRS, HDR, or EDGE in second generation systems. Third generation cellular systems harmonize voice and data access because they have been designed with Internet access in mind.

This chapter presents the fundamentals of digital communication. Many topics of relevance to digital communication are treated elsewhere in this encyclopedia. A more thorough description of wireline communication media, wireless media, and radio-frequency communication are available in other chapters of this encyclopedia. Readers who are interested in only a cursory overview can read the

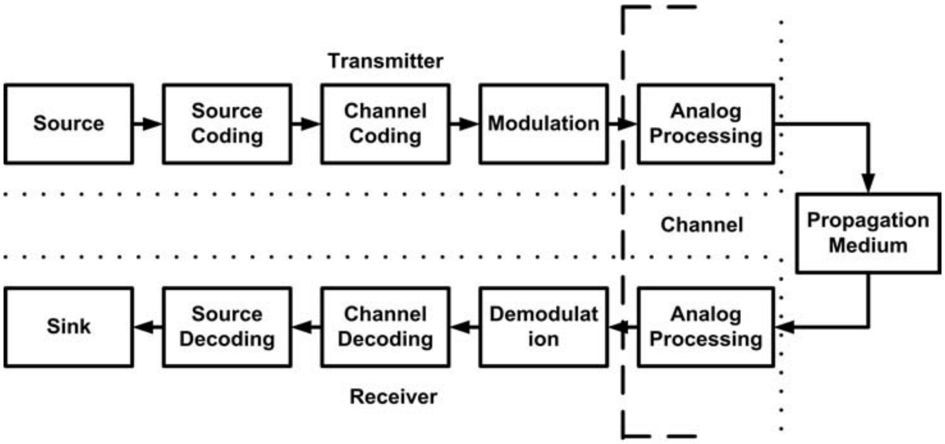


Figure 1: The components of a typical digital communication system.

section Fundamentals of Digital Communication. Those who want a more thorough treatment beyond this chapter should read Important Concepts in Digital Communication as well as consult selected references as described in the conclusion.

FUNDAMENTALS OF DIGITAL COMMUNICATION

Digital Communication System Overview

A typical digital communication system is illustrated in Figure 1. There are several steps in the transmission and reception process, some of which involve digital signal processing (DSP) and some of which involve analog processing. So it is important to distinguish that while digital communication involves the transmission of digital information, the transmission and reception process involves both digital and analog processing. In this chapter we focus on the digital signal processing aspects of digital communication.

The block diagram for a typical digital communication system in Figure 1 is divided into three parts: the transmitter, the channel, and the receiver. The transmitter processes a bit stream of data for transmission over a physical medium. The channel is the physical medium, which adds noise to and distorts the transmitted signal. It accounts for the propagation medium as well as any analog effects in the transmitter and receiver. The receiver attempts to extract the transmitted bit stream from the received signal.

The first basic transmitter block is devoted to source encoding. The purpose of source encoding is to compress the data by removing inherent redundancies. The input to the source encoder will be called $s[n]$, the source sequence. The output of the source encoder will be called $i[n]$, the information sequence. Source encoding includes both lossy and lossless compression. In lossy compression, some degradation is allowed in order to reduce the amount of data that need to be transmitted. In lossless compression, redundancy is removed, but when the encoding algorithm is inverted the signal is exactly the same. In other words, if f and g are the source encoding and decoding processes, then $\hat{s}[n] = g(i[n]) = g(f(s[n]))$, for lossy compression $s[n] \cong \hat{s}[n]$, and for lossless compression $s[n] = \hat{s}[n]$. Data compression is treated in more detail in the chapter *Data Compression*. So, from the source encoder, $s[n]$ is transformed into $i[n]$, both of which are in bits. The bit rate R_b is the rate at which information bits are transmitted through the channel.

The next block is the channel coder. Channel coding adds redundancy to the information sequence $i[n]$ in a controlled way to provide resilience to channel distortions and to improve overall throughput. Using common coding notation, for every k input bits, or information bits, there is an additional redundancy of r bits. The total number of bits is $n = k + r$; the coding rate is defined as k/n. Two types of channel codes are prevalent: forward error correction codes and error detection codes. Forward error correction codes are used to provide redundancy that enables errors to be corrected at the receiver. They come in varieties such as trellis codes, convolutional codes, and block codes (Bossert, 1999). Error detection codes, CRC (Cyclic Redundancy Check) codes being the most common, provide redundancy that allows the receiver to determine if an error occurred during transmission. The receiver can use this information either to discard the data in error or to request a retransmission.

Following channel coding, the bits are mapped to *waveforms* by the modulator. This is the demarcation point where basic transmitter-side digital signal processing for communication ends. Typically the bits are mapped in groups to *symbols*. Following the symbol mapping, the modulator converts the digital symbols into corresponding analog waveforms for transmission over the physical link. This can be accomplished by sending the digital signal through a digital-to-analog (D/A) converter into a shaping filter and, if needed, mixing it with a higher frequency carrier. Symbols are sent at a rate of R_s symbols per second, also known as the baud rate; the symbol period, $T_s = 1/R_s$, is the time difference between successive symbols.

The signal generated by the transmitter travels through a propagation medium, which could be a radio wave through a wireless environment, a current through a telephone wire, or an optical signal through a fiber, to the receiver.

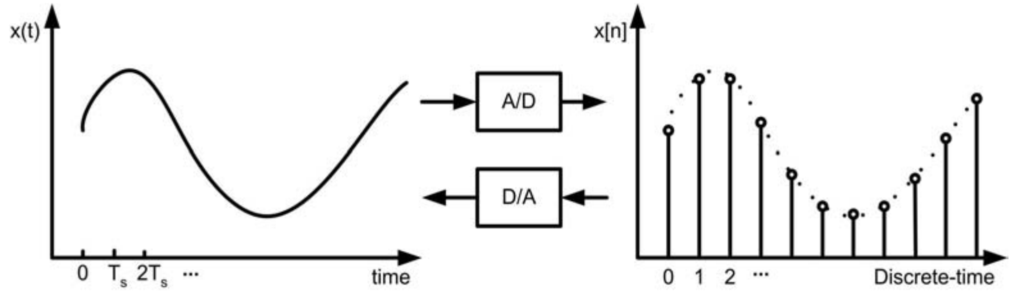


Figure 2: The relationship between a continuous-time signal and a discrete-time signal.

The first block at the receiver is the analog front end (AFE), which, at least, consists of filters to remove unwanted noise, oscillators for timing, and analog-to-digital (A/D) converters to convert the data into the digital regime. There may be additional analog components such as analog gain control and automatic frequency control. This is the demarcation point for the beginning of the receiver-side digital signal processing for digital communication.

The channel, as illustrated in Figure 1, is the component of the communication system that accounts for all the noise and intersymbol interference introduced by the analog processing blocks and the propagation medium. Noise is a random disturbance that degrades the received signal. Sources of noise include the thermal noise that results from the material properties of the receiver, the quantization noise caused by the D/A and the A/D, and the external interference from other communication channels. Intersymbol interference is a form of signal distortion that causes the transmitted signal to interfere with itself. Sources of intersymbol interference include the distortion introduced by the analog filters as well as by the propagation medium.

The first digital communication block at the receiver is the demodulator. The demodulator uses a sampled version of the received waveform, and perhaps knowledge of the channel, to infer the transmitted symbol. The process of demodulation may include equalization, sequence detection, or other advanced algorithms to help in combatting channel distortions.

The demodulator is followed by the decoder. The decoder essentially uses the redundancy introduced by the channel coder to remove errors generated by the demodulation block. The decoder may work jointly with the demodulator to improve performance or may simply operate on the output of the demodulator. Overall, the effect of the demodulator and the decoder is to produce the closest possible $\hat{\imath}[n]$ given the observations at the receiver.

The final digital block is the source decoder, which essentially reinflates the data back to the form in which they were sent: $\hat{s}[n] = g(\hat{\imath}[n])$. This is basically the inverse operation of the source encoder. After source decoding, the digital data are delivered to higher level communication protocols that are beyond the scope of the chapter.

For Internet traffic, common transmitters/receiver pairs include digital subscriber line modems, fiber optic transceivers, local area networks, and even storage devices like disk drives. While their physical media are diverse and the speeds at which they transmit may be significantly different, the fundamental model for each of these digital communication systems is the same.

Processing in the Digital Domain

There are three basic classes of signals in digital communication: *continuous-time, discrete-time,* and *digital.* Continuous-time signals are those whose value at time t is $x(t)$, where t and $x(t)$ can take values on a continuum—for instance, the real number line or the complex plane. Discrete-time signals take values only at integer times n, but the signal $x[n]$ takes values on a continuum. Finally, digital signals are those that have values at integer times and take on values on some finite (perhaps countably infinite) set.

The link between the analog and the digital domains is through the D/A and the A/D, as illustrated in Figure 2. At the transmitter, the D/A converts a digital signal $x[n]$ to an analog signal $x(t)$ essentially by letting $x(nT_s) = x[n]$ and interpolating the remaining values. At the receiver, the A/D samples the received signal $y(t)$ at some period T (typically a fraction of T_s) to produce $y[n] \cong y(nT)$, where the approximation arises because $y[n]$ is quantized to some set of values. This new signal is called $y_d[n]$ and is the digital waveform derived from the continuous-time waveform $y(t)$.

The Nyquist sampling theorem gives flexibility in the choice of whether to process $y(t)$ or $y[n]$. Ignoring quantization noise, the Nyquist sampling theorem states that if the inverse of the sampling rate is greater than twice the maximum frequency in $y(t)$ (or equivalently the bandwidth), then there is no loss in the sampling process. This implies that any processing done on the continuous-time waveform can be done equally well on the sampled waveform given that the conditions stated in the Nyquist sampling theorem are satisfied. Several practical considerations, however, make digital the domain of choice.

The biggest advantage of doing as much processing as possible in the digital domain is that it allows full exploitation of the benefits of digital technology. Many digital platforms exist that are highly customizable and easy to alter—for instance, field programmable gate arrays (FPGAs) and digital signal processors. This hardware has been designed to have very good tolerances and reproducibility. As the DSP machinery is easy to change, it is also well suited to applications where parameters may need to be adjusted over time. Analog circuitry, on the other hand, can be extremely bulky and expensive to produce. Moreover, if a change has to be made to the design,

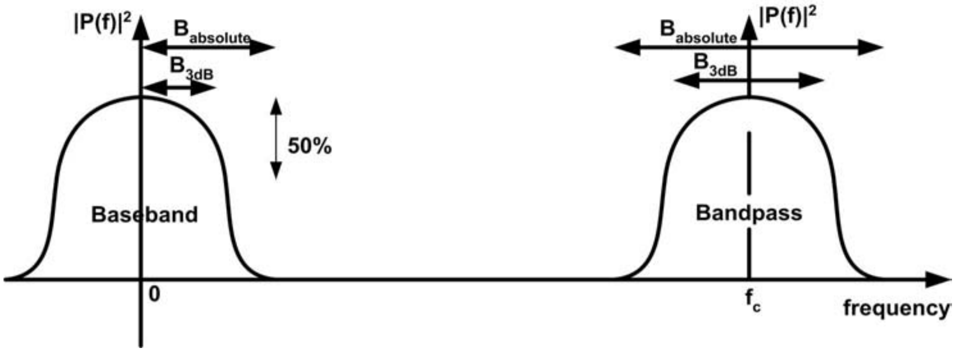


Figure 3: Illustration of two different notions of bandwidth in baseband and bandpass signals.

analog equipment may need to be redesigned as well. For these reasons, processing in the digital domain has become very relevant to digital communication. This does not mean that all analog processing can be obviated. For instance, the fundamental noise limits for a modem may be the noise produced by the analog components or the A/D converter. Another advantage of processing in the digital domain is a byproduct of Moore's law. The shrinking of transistor size offers both dramatic increases in processing power and significant reductions in overall cost. Thus, doing the majority of the processing in the digital domain improves cost, performance, and flexibility.

Key Resources: Power and Bandwidth

The two primary resources in any communication system, both digital and analog, are power and bandwidth. Systems whose performance is limited by the available power are power-limited while those that are limited by bandwidth are bandwidth-limited. Most practical systems are to some extent both power- and bandwidth-limited.

The power of a signal is roughly defined as the average energy over time. Mathematically, this is often written as $P = \lim_{T\to\infty}(1/T)\int_{-T/2}^{T/2}|x(t)|^2dt$. Power may be measured in watts but is more often measured in decibels relative to one watt (dB) or one milliwatt (dBm). The decibel is a relative measure that is defined as $(P/Q)_{dB} = 10\log_{10}(P/Q)$. When used to measure the power of P in dB, Q is assumed to be 1 watt, while to measure the power of P in dBm, Q is assumed to be 1 milliwatt.

There are two different but related notions of power in a communication system: transmitted power and received power. Naturally, the transmitted power is the average energy over time of the transmitted signal while the received power is the average energy over time of the received signal.

The transmitted power in a communication system is limited by the maximum power available to transmit a signal. Generally system performance is better if there is high transmitted power (and thus received power). Practical constraints on cost, properties of the transmission medium, battery life (in mobile systems), or regulatory constraints, generally motivate having a low transmitted power.

Since propagation media are lossy and dispersive, the received power is a function of the transmitted power and the channel. In all media the loss due to the channel increases as some function of the distance between the transmitter and receiver. Thus the larger the distance, the smaller the received power.

The minimum received power required at the receiver, known as the *receiver sensitivity,* is determined by the parameters of the system, the quality of the hardware, as well as the desired operating characteristics. The range of the system can be inferred from the ratio of the maximum transmitted power to the minimum received power. Generally, increased data rate at a given bandwidth or lower bit error rates increases the required minimum received power.

Besides power, *bandwidth* is the other design constraint in a communication system. Unfortunately, there are many definitions of *bandwidth* and different notions are used in different systems. The most generic definition of the bandwidth of a signal $x(t)$ is the portion of the frequency spectrum $X(f) = \int_{-\infty}^{\infty} x(t)e^{-j2\pi ft}dt$ for which $X(f)$ is nonzero. Since the true bandwidth of a finite duration signal is infinite, systems often use the "3 dB bandwidth," which is the contiguous range of frequencies over which the power spectrum is at least 50% of the maximum value. Other definitions of bandwidth are also possible; see Couch (2001) for details.

It is important to note that the definition of bandwidth differs depending on whether the communication system is *baseband* or *bandpass.* Baseband communication systems operate at DC while bandpass communication systems convey information at some carrier frequency f_c. Figure 3 illustrates the different bandwidth notions of absolute and 3dB bandwidth in each of these types of systems.

The bandwidth available in a communication system depends on the transmission medium. In wireless systems, bandwidth is a precious and expensive commodity that is regulated by the government. Thus while the theoretical bandwidth of the wireless medium is significant, only a fraction is available to a given communication system. In wireline systems, the bandwidth is determined by the type and quality of the cable and the interconnects. Generally, the larger the bandwidth, the larger the potential data rate that can be supported. Exploiting larger bandwidths, however, typically requires more sophisticated receiver processing algorithms and more expensive analog components.

Bandwidth and power are related through the concept of the power spectral density (PSD). The PSD is a measure of the power in a signal as a function of frequency. The integral of the PSD is hence the power of the signal.

Measures of Performance

There are a number of potential measures that are used to evaluate the performance of a digital communication system. The choice of a performance measure depends significantly on the application and other aspects of the system. Broad classes of performance measures include the probability of error, the outage probability, and the capacity. In this chapter we will discuss the probability of error and capacity. The probability of error is a measure of the rate at which errors occur, while the capacity is a measure of the maximum data rate that can be supported by a channel with a given SNR and an arbitrarily small probability of error.

Of the two measures, the probability of error is the more pragmatic indicator of system performance. There are various flavors of the probability of error including the probability of bit error, the probability of symbol error, and the probability of frame error. Equivalently, these measures are known as the bit error rate (BER), the symbol error rate (SER), and the frame error rate (FER). Often the SER and the FER can be determined from the BER; therefore the BER is the most common performance metric.

Essentially the BER provides the average number of bit errors. For example, a BER of 10^{-2} means that on an average about one bit out of every one hundred will be in error. The BER can be measured at various places in the receiver but is typically most meaningful after demodulation (the uncoded BER) and after error correction (the coded BER).

The performance needs of the application determine the required BER. For example, voice communication in cellular systems might require a coded BER of 10^{-2}, while data communication in the same system might require a coded BER of 10^{-5}. In most communication systems the uncoded BER is a function of the data rate and the modulation scheme and can be readily related to the *signal-to-noise ratio* (SNR). The SNR is essentially the ratio of the received signal power to the noise power in the signal bandwidth. Thus the BER is a function of both the received power and the bandwidth of the signal though more generally the channel model also plays a role.

The fundamental limit to data communications can be most simply described by the so-called capacity of a channel, C, which is the maximum average number of bits that can be supported by a channel with a given SNR at an arbitrarily small probability of error. The capacity is measured in units of bits per second and is essentially a bound on the achievable data rate. Often the capacity is normalized by the bandwidth. The normalized capacity C/B measures the bits per channel use in units of bits per second per hertz. Unlike the BER, the capacity provides an upper bound (instead of the actual performance) of a communication system, since it is optimized over all possible modulation and coding schemes. Like the BER, the capacity is typically a function of the SNR, the bandwidth, and the channel.

The capacity is a measure for determining the fundamental limit on the data rate imposed by the given communication channel. The BER is more useful for evaluating the performance of an actual coding and modulation scheme. Typically, a target BER will be defined and a coding and modulation scheme will be proposed to achieve the largest data rate possible R. Naturally, it should be the case that $R < C$. Spectrally efficient digital communication systems have a rate R that closely approaches the capacity C for the desired operating point.

IMPORTANT CONCEPTS IN DIGITAL COMMUNICATION

Modulation

The modulator in a digital communication system maps binary data onto waveforms for transmission over the physical channel. The modulator maps a group of bits (or symbols) onto a finite number of waveforms during each symbol period. Binary modulations map each bit to one of two possible waveforms, while M-ary modulations map each group of $\log_2 M$ bits to one of M possible waveforms. The analog waveforms are designed with the constraints of the channel such as the bandwidth or the carrier frequency in mind.

There are two basic forms of modulation, namely linear modulation and nonlinear modulation. Linear modulation schemes are typically more spectrally efficient; that is, they are able to come closer to the capacity. Nonlinear modulations typically have other properties, such as constant envelope, that make them easier to implement and less susceptible to various impairments in the channel. The choice of a modulation scheme depends on the desired throughput, the target bit error rate, the spectral efficiency of the modulation scheme, the power efficiency of the modulation scheme, robustness to impairments, and the implementation cost and complexity.

Modulations may also have memory or be memoryless. When a symbol is a function of only the bits from the current symbol period, it is said to be memoryless. When a symbol is a function of the bits from previous symbol periods, the modulation is said to have memory. Having memory in the modulation scheme may have some practical advantages, such as reducing the peak-to-average ratio of the transmitted signal or simplifying noncoherent detection at the receiver. Modulations with memory can also provide some additional resilience to errors; thus they are sometimes called coded modulation schemes.

To illustrate the concept of modulation, in particular linear modulation, let us first define the concept of a signal space and then relate that to common modulation formats found in practice. A vector space is defined to be a set of vectors (in this case continuous signals), $\{\phi_i(t)\}$, with two operations: (1) addition of those vectors (i.e., $\phi_i(t) + \phi_j(t)$ is defined and is an element of the vector space) and (2) multiplications of those vectors by a scalar (i.e., $k\phi_i(t)$ is defined and is an element of the vector space). Some other technical rules for being a vector space can be found in Anton and Rorres (1991). In communications, typically these vectors are orthogonal; that is,

$$\int_{-\infty}^{\infty} \phi_i(t)\phi_j(t)dt = \delta_{ij}, \tag{1}$$

where δ_{ij} is 1 for $i = j$ and 0 otherwise. Since these waveform vectors are orthogonal, they are also a basis for the vector space, and are sometimes referred to as basis functions.

A digital communication system may be baseband or bandpass depending on whether the symbols are conveyed at baseband (DC) or at some carrier frequency (see Figure 3). ADSL (asymmetric digital subscriber line) is an example of a baseband communication system. Most digital communication systems are bandpass, including all narrowband wireless systems, optical systems, and cable modems. Though bandpass systems convey information at a carrier frequency, the modulator and demodulator do not need to generate signals at that carrier frequency. Instead, the modulator and demodulator work with the *baseband equivalent* waveform. At the transmitter, the *upconverter* in the analog processing block converts the baseband equivalent signal to the bandpass signal by shifting it to the desired carrier frequency. At the receiver, the *downconverter* in the analog processing block shifts the bandpass signal down to zero frequency. The advantage of the baseband equivalent notion is that it makes the digital operations of the communication system independent of the actual carrier frequency.

Let us first consider baseband pulse amplitude modulation (PAM). PAM transmission is used in HDSL (high bitrate digital subscriber line), ISDN, and optical transmission. It is a linear and memoryless modulation. An M-PAM system is a form of M-ary modulation in which $m = \log_2 M$ bits at a time are mapped to an element of the set of M possible amplitudes, $\mathcal{C}_{PAM}$, which is the constellation. For 4-PAM, a set of possible amplitudes is $\mathcal{C}_{PAM} = \{-3, -1, 1, 3\}$, which are equally spaced apart. A pulse-shaping filter $\phi(t)$ with a bandwidth B is modulated to produce the transmitted waveform $x(t) = \sum_n x[n]\phi(t - nT_s)$, where $x[n] \in \mathcal{C}_{PAM}$. So the set $\mathcal{C}_{PAM}$ is the constellation, $x[n]$ is the symbol transmitted starting at time nT_s, and $\phi(t)$ is the basis waveform. The spectrum of the PAM waveforms is determined by the pulse-shaping filter $\phi(t)$. The choice of a pulse-shaping filter is a complicated one involving several competing requirements, including resistance to timing jitter, minimized spectral bandwidth, and noise immunity.

A nice generalization of PAM that is preferable for bandpass systems is known as quadrature amplitude modulation (QAM). As with PAM, QAM is a linear and memoryless modulation. An M-QAM modulation is defined for M that is a power of 4. Let $m = 1/2 \log_2 M$ and consider the same set of 2^m possible amplitudes $\mathcal{C}_{PAM}$ and pulse-shaping waveform $\phi(t)$. For M-QAM, at some carrier frequency f_c, the transmitted waveform is

$$x(t) = \left(\sum_n i[n]\phi(t - nT_s)\right)\cos(2\pi f_c t) - \left(\sum_n q[n]\phi(t - nT_s)\right)\sin(2\pi f_c t), \tag{2}$$

where $i[n] \in \mathcal{C}_{PAM}$ corresponds to the symbol transmitted on the inphase component $\cos(2\pi f_c t)$ and $q[n] \in \mathcal{C}_{PAM}$ corresponds to the symbol transmitted on the quadrature component $\sin(2\pi f_c t)$. Assuming that f_c is much greater than $1/B$ the modulated inphase and quadrature components are orthogonal and thus QAM has double the data rate of a PAM system, which would use only the inphase or quadrature component. Of course, recalling the discussion about Figure 3, note that a bandpass QAM system also uses twice the bandwidth of a baseband PAM system; thus the spectral efficiency of QAM at bandpass and PAM at baseband is the same. The ordered pair $(i[n], q[n])$ is the symbol transmitted starting at time nT_s, and $\phi(t)\cos(2\pi f_c t)$ and $\phi(t)\sin(2\pi f_c t)$ are the basis waveforms, using the narrowband approximation. For M-QAM, the constellation $\mathcal{C}_{QAM}$ is composed of all possible ordered pairs that can be generated from choosing the $2^m - PAM$ points for the inphase and quadrature components. Thus the M-QAM constellation has two dimensions. The baseband equivalent of the QAM signal in (2) is a complex function and is given by

$$\sum_n (i[n] + jq[n])\,\phi(t - nT_s),$$

where $j = \sqrt{-1}$. Quadrature amplitude modulation schemes have complex baseband equivalents to account for both the inphase and quadrature components. The spectrum of the QAM waveform is determined by the pulse-shaping filter $\phi(t)$ and is shifted in frequency by f_c. PAM and QAM modulation are illustrated in Figure 4.

Not all bandpass systems are capable of using QAM modulation. For example, in optical transmission, phase information is not available, so the constellations are defined for positive amplitudes only. If the constellations values are 0 and A, this is known as on–off keying (OOK).

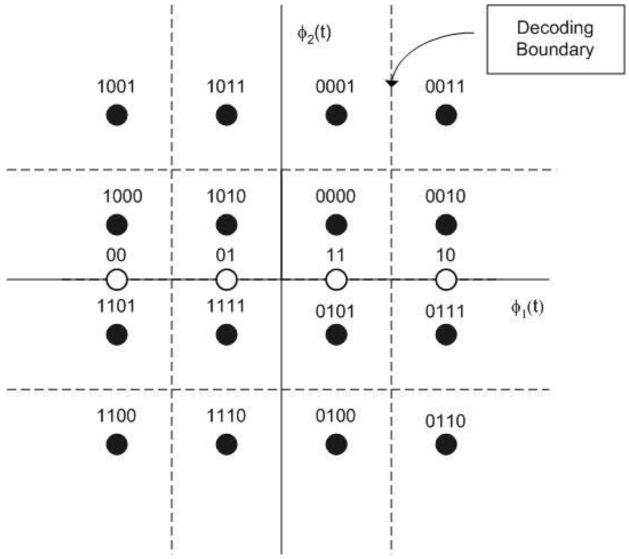


Figure 4: Signal space illustration of PAM and QAM signalling. The solid black circles are QAM and the white circles are PAM. The numbers above represent the mapping of bits to the constellation points.

Another common form of modulation is multicarrier modulation, in which several carriers are modulated simultaneously. These carriers can be thought of as a basis waveform modulated by sinusoids of differing frequencies. This is not easy to accomplish, so an IFFT (inverse fast Fourier transform) is used to approximate this operation. The IFFT of a set of constellation points is in fact the sum of a set of sampled sinusoids at differing frequencies multiplied by the constellation points. In the digital domain, each of these waveforms can be independently demodulated. These samples are sent through a single pulse-shaping filter (basis waveform) and transmitted through the channel. The receiver samples the waveform and performs the inverse operation, an FFT. Multicarrier modulation has become important because of its robustness to impulsive noise, its ease of equalization, and its ability to do spectral shaping by independently controlling the carriers. Discrete multitone (DMT) is the most common baseband version of multicarrier modulation, while orthogonal frequency division multiplexing (OFDM) is the most common bandpass version.

Intersymbol Interference Channels

After modulation, the analog waveform $x(t)$ corresponding to an input sequence of bits is transmitted over the communication medium. Communication media, whether fiber optic cable, coax cable, telephone cable, or free space, generally have a dispersive effect on the transmitted signal.The effect of the medium on $x(t)$ is often modeled using a concept from signal processing known as a linear and time-invariant (LTI) system. The linearity implies that if $y_k(t)$ is the response to $x_k(t)$ then $\alpha y_1(t) + \beta y_2(t)$ is the response to the input signal $\alpha x_1(t) + \beta x_2(t)$. The time-invariance means that the response to $x(t+\tau)$ is $y(t+\tau)$; that is, the behavior of the channel is not a function of time. Practically all physical channels are time-varying (due to changes in environmental factors), especially wireless channels; however, over short periods of time, they can be modeled as time-invariant. Optical channels can also exhibit nonlinear behavior and thus other models may sometimes be appropriate.

The LTI assumption about the communication medium allows the distortion to the input signal to be modeled using the convolution operation

$$y(t) = \int_{-\infty}^{\infty} h(\tau)x(t-\tau)d\tau.$$

The function $h(\tau)$ is known as the impulse response of the channel and includes all the analog effects such as filtering at the transmitter and receiver, in addition to the distortion in the medium. From basic Fourier transform theory, LTI systems have the nice property that in the frequency domain $Y(f) = H(f)X(f)$, where $Y(f)$ is the Fourier transform of $y(t)$, $H(f)$ is the Fourier transform of $h(t)$, and $X(f)$ is the Fourier transform of $x(t)$. Essentially, the channel acts as a frequency-selective filter that operates on the input signal. LTI systems induce distortion that is multiplicative in the frequency domain.

In an ideal channel, $|H(f)| = 1$ and there is no distortion (only a delay) of the input signal. Equivalently, in the time domain an ideal channel produces $y(t) = ce^{jd}x(t-\tau)$, where τ is an arbitrary delay. When the channel is not ideal, a more serious problem known as intersymbol interference (ISI) is encountered. To illustrate this concept, suppose that $x(t)$ is generated at baseband using PAM as described in the previous section.

In the absence of a channel, ideal sampling of $y(t)$ at the receiver at time mT_s (more details on this in the next section) results in $y[m] = \sum_n x[n]\phi(mT_s - nT_s)$. The pulse-shaping filter $\phi(t)$, however, is often a Nyquist pulseshape, which means that $\phi(0) = 1$ and $\phi(nT_s) = 0$ for $n \neq 0$. Thus sampling at time mT_s yields the symbol $x[m]$. Now consider a nonideal channel. Let $\tilde{\phi}(t)$ be the convolution of $\phi(t)$ and $h(t)$. For a nontrivial channel, it will generally be the case that $\tilde{\phi}(t)$ is no longer a Nyquist pulseshape and thus $y[m] = \sum_n x[n]\tilde{\phi}((m-n)T_s)$. In this case there are multiple superpositions of symbols at each sampling instant—this leads to intersymbol interference. Compensation for intersymbol interference is known as equalization and is an important part of the receiver processing when ISI is present.

Noise and Interference

Noise and interference are the most ubiquitous forms of degradation in any communication system. Essentially, both can be modeled as random disturbances that are unrelated to the desired signal. Intersymbol interference also results in degradation; however, the effect is different because it causes the signal to interfere with itself. Noise usually refers to the disturbances generated in the receiver as a result of the analog components, analog-to-digital conversion, and material properties of the receiver. Generally, noise can be reduced by using higher quality materials but never eliminated. Interference usually refers to disturbances generated by external signals. Typically interference has more signal structure than noise and thus it can be mitigated by more complex processing at the expense of higher cost.

There are various sources of noise in communication systems. Common examples include thermal noise, shot noise, and quantization noise. Thermal noise is a result of the Brownian random motion of thermally excited electrons. It is generated by resistors and the resistive parts of other devices such as transistors. Shot noise is more impulsive and may be more related to the signal, for example, the random arrival rates of photons in optical systems. Quantization noise is a result of digitizing the amplitude of the discrete-time signal and is often modeled as another source of thermal noise. We will focus our explanation on thermal noise.

Because noise is fundamentally not deterministic, it is often modeled as a random process. For thermal noise, the Gaussian random process has been found to be adequate for the job. When used to model thermal noise, the process is assumed to be zero mean, uncorrelated from sample to sample and to have a variance σ^2 that is generally proportional to kBT_e, where k is Boltzmann's constant ($1.23 \times 10^{-23} J/K$), B is the signal bandwidth, and T_e is the effective noise temperature of the device. T_e is a parameter determined by the analog portion of the receiver. For thermal noise, the variance increases linearly as a function of the bandwidth. Thus signals with larger bandwidths incur an additional noise penalty, while at the same time enjoying a higher signaling rate.

The effect of thermal noise is additive; therefore the received signal can be written $z(t) = y(t) + v(t)$, where $v(t)$ is a realization of the noise process. Since $v(t)$ is unknown to the receiver, its presence degrades the performance of subsequent processing blocks. The severity of thermal noise is quantified by the SNR.

The origin of interference is usually an undesired communication signal. Examples include adjacent channel interference, crosstalk, and co-channel interference. Adjacent channel interference refers to the interference caused by signals operating in adjacent frequency bands. Since practical signals cannot have a finite absolute bandwidth, when the carrier frequencies of two different signals are close to each other there is often leakage from one signal to the other. Crosstalk is a form of interference in wireline systems. It results from electromagnetic coupling among the multiple twisted pairs making up a phone cable. Co-channel interference is the wireless equivalent of crosstalk. Because of the limited availability of frequencies, wireless cellular systems reuse each carrier frequency. Essentially, co-channel interference is the interference among users sharing the same communication frequency.

Like thermal noise, interference is also additive. Thus a signal with interference may be written $z(t) = y(t) + \sum_k y_k(t)$, where $y_k(t)$ refers to the distorted interfering signals. Performance degrades because $y_k(t)$ is both random and unknown at the receiver. More generally, noise is also present; thus

$$z(t) = y(t) + \sum_k y_k(t) + v(t).$$

In some cases, the interference is modeled as another Gaussian noise source. Then performance is characterized by the signal to interference plus noise (SINR) ratio, which is essentially $P_y/(P_i + P_v)$, where P_y is the power in the desired signal, P_i is the power in the sum of the interfering signals, and P_v is the power in the noise. Systems for which $P_i \ll P_v$ are called noise-limited, while those for which $P_v \ll P_i$ are called interference-limited. If the source of the interference is crosstalk or co-channel interference, then the interfering signals $\{y_k(t)\}$ all have a structure (modulation, coding, etc.) similar to that of the desired signal. In this case, advanced signal processing algorithms can be used to mitigate the impact of the interference. Examples of algorithms include joint demodulation, where all the signals are demodulated simultaneously, interference cancellation, where the interference is partially cancelled, and optimum filtering, where filters are constructed that partially eliminate the interference. Removing or mitigating interference improves performance by reducing the required transmitted power to achieve a given BER for a given data rate or by allowing the data rate at a given BER to be increased for a given transmitted power.

Timing and Synchronization

Prior to demodulation and symbol recovery at the receiver a number of timing and synchronization tasks need to be performed, including phase synchronization, frequency synchronization, symbol timing, and frame synchronization. Synchronization is performed to ensure that the transmitter and receiver operate in a synchronous manner. The process of synchronization typically involves estimation of the synchronization error, followed by its correction. Typically the processing required for synchronization is done in a mixture of the analog and digital domains.

In bandpass systems, information is modulated onto sinusoids as illustrated in the QAM example in (2). To demodulate this signal at the receiver, these sinusoids must be exactly reproduced. The problem of ensuring that the phases are accurate is known as phase synchronization. The problem of estimating the transmitted carrier frequency is known as frequency synchronization. Estimating and tracking the phase of the sinusoid is typically more difficult than the frequency; however, phase differences can sometimes be included as part of the channel and removed during equalization.

At the receiver, the A/D samples the analog waveform for subsequent digital processing. Optimal processing requires two aspects of symbol synchronization: symbol timing and sampling clock recovery. Symbol timing is knowing exactly where to sample the received signal. Even in systems with ideal channels, symbol timing errors can lead to intersymbol interference. Often the symbol timing problem is solved by oversampling the received signal and choosing the best subsample. Sampling clock recovery refers to ensuring that the sampling period T_s at the receiver is identical to that at the transmitter. Sampling clock recovery is typically more important in baseband systems because in bandpass systems the sampling clock can be derived from the carrier.

In systems where the fundamental unit of information is a frame and not a symbol, an additional synchronization step is required. This process, known as frame synchronization, is required to determine where the beginning of the frame is located. Frame synchronization is often assisted by the presence of synchronization sequences that mark the beginning of the frame.

Demodulation

The goal of the demodulator is to convert the sampled received waveform back into a sequence of bits. Demodulation is obviously highly dependent on the modulation and the channel; therefore this section provides only a cursory overview.

The first step in the demodulation process is to sample the waveform. Typically, this is done using a front end filter that filters out unwanted noise, followed by a sampler. The sampled data come from an A/D converter, so the data are in the digital domain. The sampled signal includes residual noise left after filtering, the noise from the sampling device, and the signal of interest. For example, assuming perfect timing, synchronization, and no interference, in the presence of an ideal channel the sampled PAM signal at the receiver $y(nT_s)$ is

$$y[n] = x[n] + v[n],$$

where $x[n]$ is the transmitted PAM symbol and $v[n]$ represents the sampled thermal noise and the quantization noise. The samples $y[n]$ are sent through a decision device that processes the data to make a decision. Typically, the decision device determines the most likely symbol from the given constellation that was transmitted. An inverse symbol mapping operation then converts the symbols to bit form.

Because the noise is unknown to the receiver, the role of the decision device is to produce its best "guess" about the transmitted data. One common criterion is to find the symbol that is the most likely input given the observations. If X and Y are vectors representing the input $x[k]$ and the output $y[k]$, respectively, then

$$\hat{X} = \arg\max_{X} P(X|Y),$$

where $P(X|Y)$ is the probability that $x[k] = X$ given the observation of $Y = y[k]$. The maximization is taken over all possible points in the constellation to find the point with the maximum conditional probability and is called the maximum a posteriori decision. When the source data are equally likely, it turns out that this is equivalent to the maximum likelihood detection rule, which is given by

$$\hat{X} = \arg\max_{X} P(Y|X).$$

In this case the decision rule determines the symbol that was most likely to have produced the observation. For the additive white Gaussian noise channel (AWGN), the above conditional probabilities have a known form. In the AWGN channel there is no intersymbol interference and the only source of degradation is uncorrelated Gaussian noise. For instance, when the input symbols are equally likely, it turns out that the detection principle is simply to minimize the Euclidean distance between the observation and the set of possible inputs,

$$\hat{X}[k] = \arg\min_{x[k]\in C} ||y[k] - x[k]||^2. \tag{3}$$

In this case, the detector is known as a slicer. The operation of the slicer can be described using Figure 4. For a QAM/PAM waveform, if a received sample is within a decoding boundary of a point, it is mapped to that point. Because of the simple decoding boundaries, the test is essentially a series of threshold tests; hence the name slicer.

In the absence of an ideal channel, even with perfect timing and synchronization, there will be intersymbol interference and thus the sampled PAM signal may have the form

$$y[n] = \sum_{l=0}^{L} h[l]x[n-l] + v[n],$$

where $h[l]$ is the sampled equivalent channel impulse response. Optimum decoding requires considering the channel response in the decision device. Due to the memory in the channel it is no longer possible to make a decision on a symbol-by-symbol basis. Instead, sequences must be decoded. Thus given a sequence of observations $\{y[p]\}_{p=0}^{P-1}$ we must determine the sequence $\{x[n]\}_{n=0}^{N-1}$ that was most likely to have been transmitted. We allow for $P \geq N$ at the receiver to account for multiple observations of the received signal, via oversampling or multiple antennas, for example. Clearly the complexity of the search grows with both N and P. Using the fact that the memory of the channel is finite, however, allows lower complexity approaches such as the Bahl, Cocke, Jelenick, and Raviv (BCJR) algorithm to help in maximum a posteriori decoding and the Viterbi algorithm for maximum likelihood decoding (see, e.g., Wither & Kim, 2002, for details).

Alternatively, to correct for intersymbol interference, many transmission systems use equalizers that attempt to remove the effect of the channel prior to the slicing operation. Some common equalizers are zero-forcing equalizers (ZFE) that invert the channel, minimum mean squared error (MMSE) equalizers that include the effects of noise, and decision feedback equalizers (DFE) that use the detected symbols to remove some portion of the trailing intersymbol interference. Equalization generally gives inferior performance relative to sequence decoding but offers much lower complexity.

A PERFORMANCE EXAMPLE

The AWGN channel provides an analytically tractable baseline case to which performance in other channels can be compared. First consider the capacity of the AWGN channel. It can be shown in this case that the capacity expression is remarkably simple,

$$C = B\log_2(1 + SNR), \tag{4}$$

where B is the channel bandwidth. The theoretical spectral efficiency that can be achieved in this channel is thus C/B bits per second per Hertz. The normalized capacity as a function of SNR is illustrated in Figure 5.

The capacity expression in (4) provides an interesting means of evaluating system performance. For instance, if a coded system provides 3 dB increased immunity to noise, the SNR will increase by a multiple of 2 ($10\log_{10} 2 \cong 3\,\mathrm{DB}$). Hence the amount of information that can be transmitted will increase by about 1 bit per transmission

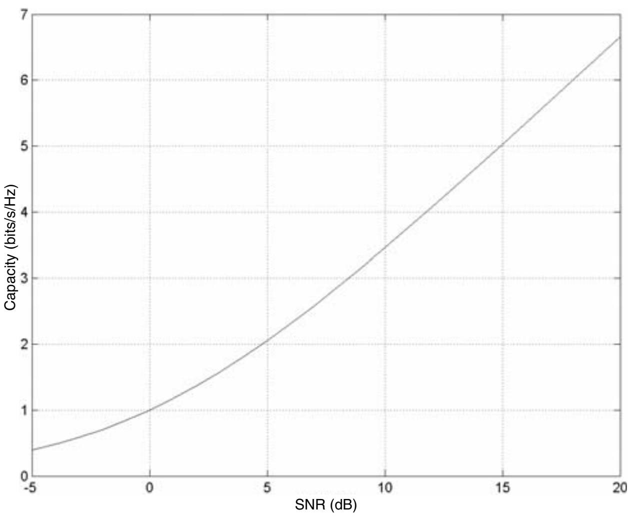


Figure 5: The normalized capacity of an AWGN channel as a function of the SNR.

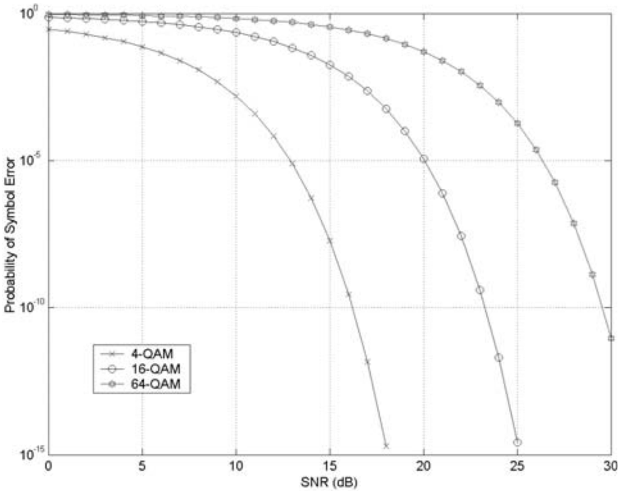


Figure 6: The symbol error rate for QAM transmission in an AWGN channel as a function of SNR.

for high SNR because $\log_2(1 + 2 \cdot \text{SNR}) \cong \log_2(2 \cdot \text{SNR}) \cong \log_2 2(\text{SNR}) + 1$.

Unlike the capacity, the symbol error rate in an AWGN channel is a function of the modulation scheme that is employed. For uncoded M-PAM or M-QAM transmission, the probability of symbol error is given by

$$P_e = 2\left(1 - \frac{1}{M}\right) Q\left(\sqrt{\frac{3\,SNR}{M^2 - 1}}\right) \quad \text{for PAM} \qquad (5)$$

$$= 4\left(1 - \frac{1}{\sqrt{M}}\right) Q\left(\sqrt{\frac{3\,SNR}{M - 1}}\right)$$

$$-4\left(1 - \frac{1}{\sqrt{M}}\right)^2 \left(Q\left(\sqrt{\frac{3\,SNR}{M - 1}}\right)\right)^2$$

$$\text{for QAM,} \qquad (6)$$

where $Q(x) = 1/\sqrt{2\pi} \int_x^\infty e^{-t^2/2} dt$ for $x \geq 0$ is the area under the tail of the Gaussian probability distribution function. The probability of symbol error for QAM transmission is illustrated in Figure 6 as a function of SNR. Notice how the error rate is exponentially decreasing as the SNR increases. For a given probability of error (at high SNR), observe that there is approximately a 6 dB difference between the SNR required for 4-QAM and 16-QAM or between 16-QAM and 64-QAM.

By inverting the formulas in (5) and (6), for a target probability of error, a useful expression for the maximum spectral efficiency obtained is

$$R = \log_2\left(1 + \frac{SNR}{\Gamma}\right), \qquad (7)$$

where the gap, Γ, can be calculated as a function of M and the target probability of error. Conveniently, (7) allows direct comparison with the capacity in (4). The gap, Γ, effectively determines the loss in capacity, i.e., the difference between the actual spectral efficiency and the maximum spectral efficiency. Coded modulation schemes generally reduce the gap (towards the ultimate limit of $\Gamma = 1$). The effect of coding is most often expressed in dB as the coding gain, ϕ_{dB}, making the effective gap smaller, so that $\Gamma_{db}^{\text{new}} = \Gamma_{dB} - \phi_{dB}$.

CONCLUSION: ADDITIONAL RESOURCES

Digital communications is a broad area that draws on many different but related aspects of electrical engineering. Perhaps the standard academic references for digital communication are *Digital Communications* by John G. Proakis and *Digital Communication* by Edward Lee and David Messerschmitt. The text *Digital Communications* by Bernard Sklar provides an intuitive presentation of the concepts of digital communications. A good online manuscript is John Cioffi's. A good technical discussion of digital communication in wireless systems is found in *Principles of Mobile Communication* by Gordon L. Stuber. The standard reference for digital signal processing is *Discrete-Time Signal Processing* by Alan V. Oppenheim and Ronald W. Schafer. For a basic reference on vector spaces, *Elementary Linear Algebra: Applications Version* by Howard Anton and Chris Rorres is a good text.

There are many advanced concepts in digital communication that were just barely covered. For example, *Elements of Information Theory* by Thomas M. Cover and Joy A. Thomas provides a more thorough introduction to information theory. The book *Synchronization Techniques for Digital Receivers* by Umberto Mengali and Aldo N. D'Andrea provides a current treatment of synchronization in digital communication systems. Forward error correction is a topic that was only briefly mentioned yet is of significant importance. A classic reference is *Error Control Coding: Fundamentals and Applications* by Shu Lin and Dan J. Costello. A text that treats some current topics is *Fundamentals of Codes, Graphs, and Iterative Decoding* by Stephen B. Wicker and Saejoon Kim.

Current research in digital communication appears in a variety of journals including the *IEEE Transactions on Communications*, the *IEEE Transactions on Signal Processing*, and the *IEEE Transactions on Information Theory*, among others.

GLOSSARY

A/D Analog-to-digital converter.

AFE Analog front end.

Bandpass A type of communication signal that has information modulated onto a carrier.

Baseband A type of communication signal that does not have information modulated onto a carrier.

BER Bit error rate.

Capacity The maximum data rate at which nearly errorless transmission can occur.

Carrier Name for the high-frequency sinusoid that shifts the spectrum of a baseband signal to higher frequencies, making it bandpass.

Channel Coding The process of adding redundancy to a transmitted data stream for the purpose of improving resilience to errors caused by the channel.

CDPD Cellular digital packet data.

D/A Digital-to-analog converter.

Demodulation The process of extracting transmitted digital data from the continuous waveform observed at a receiver.

Downconversion Process of converting a bandpass signal to a baseband signal by removing the carrier frequency.

DSP Digital signal processing.

EDGE Enhanced data rates for GSM evolution.

FPGA Field programmable gate array.

HDR High data rate.

GPRS General packet radio service.

Modulation The process of converting digital data to continuous waveforms for transmission on a communication medium.

PAM Pulse amplitude modulation.

PSD Power spectral density is a measure of the power in a signal as a function of frequency.

QAM Quadrature amplitude modulation.

Receiver Sensitivity The minimum required signal level for a receiver to be able to demodulate a received signal at the desired quality.

SNR Signal to noise ratio. Essentially the ratio of the received signal power to the noise power.

Source Generic name for the component that generates the information stream which is the input to a transmitter.

Source Encoding The process of removing redundancy from the information stream provided by a source.

Symbol A representation of a set of bits in the digital or analog domain.

Synchronization The process of ensuring that a transmitter and receiver operate in a synchronous manner.

Upconversion Process of converting a baseband signal to a bandpass signal by increasing the carrier frequency.

CROSS REFERENCES

See *Conducted Communications Media; Data Compression; Propagation Characteristics of Wireless Channels; Radio Frequency and Wireless Communications; Wireless Communications Applications.*

REFERENCES

Bossert, M. (1999). *Channel Coding for Telecommunications.* New York: Wiley.

Couch, L. W., II (2001). *Digital and Analog Communication Systems* (6th ed.). Upper Saddle River, NJ: Prentice Hall.

Anton, H., & Rorres, C. (1991). *Elementary Linear Algebra: Applications Version* (6th ed.). New York: Wiley.

Wicker, S. B., & Kim, S. (2002). *Fundamentals of Codes, Graphs, and Iterative Decoding.* Boston: Kluwer Academic.

Proakis, J. G. (2000). *Digital Communications* (4th ed.). Boston: McGraw–Hill.

Lee, E., & Messerschmitt, D. (1994). *Digital Communication* (2nd ed.). Boston: Kluwer Academic.

Sklar, B. (2000). *Digital Communications* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Cioffi, J. M. EE 379A—Digital Communications: Fundamentals and Signal Processing Applications. Retrieved from http://www.stanford.edu/class/ee379a/

Cioffi, J. M. EE379B—Digital Communication II: Coding. Retrieved from http://www.stanford.edu/class/ee379b/

Cioffi, J. M. EE 379C—Advanced Digital Communication. Retrieved from http://www.stanford.edu/class/ee379c/

Stuber, G. L. (2001). *Principles of Mobile Communication* (2nd ed.). Boston: Kluwer Academic.

Oppenheim, A. V., Schafer, R. W., & Buck, J. R. (1999). *Discrete-Time Signal Processing* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Cover, T. M., & Thomas, J. A. (1991). *Elements of Information Theory.* New York: Wiley–Interscience.

Mengali, U., & D'Andrea, A. N. (1997). *Synchronization Techniques for Digital Receivers.* New York: Plenum.

Costello, D. J., & Lin, S. (1983). *Error Control Coding.* Englewood Cliffs, NJ: Prentice Hall.

IEEE Transactions on Communications. Retrieved May 6, 2003, from http://www.comsoc.org/pubs/jrnal/transcom.html

IEEE Information Theory Society Homepage. Retrieved May 6, 2003, from http://www.itsoc.org

IEEE Signal Processing Society Homepage. Retrieved May 6, 2003, from, http://www.ieee.org/organizations/society/sp/

FURTHER READING

Proakis, J. G. (2000). *Digital Communications* (4th ed). Boston: McGraw–Hill.

Lee, Edward, & Messerschmitt, David (1994). *Digital Communication* (2nd ed.). Boston: Kluwer Academic.

Sklar, B. (2000). *Digital Communications* (2nd. ed.). Upper Saddle River, NJ, Prentice Hall.

Cioffi, J. M., EE 379A—Digital Communications: Fundamentals and Applications. Retrieved from http://www.stanford.edu/class/ee379a/

Cioffi, J. M., EE379B—Digital Communication II: Coding. Retrieved from http://www.stanford.edu/class/ee379b/

Cioffi, J. M., EE 379C—Advanced Digital Communication. Retrieved from http://www.stanford.edu/class/ee379c/

Stuber, G. L. (2001). *Principles of Mobile Communication* (2nd ed.). Boston: Kluwer Academic.

Oppenheim, A. V., Schafer, R. W., & Buck, J. R. (1999). *Discrete-Time Signal Processing* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Cover, T. M., & Thomas, J. A. (1991). *Elements of Information Theory.* New York: Wiley–Interscience.

Mengali, U., & D'Andrea, A. N. (1997). *Synchronization Techniques for Digital Receivers.* New York: Plenum.

Costello, D. J., & Lin, S. (1983). *Error Control Coding.* Eaglewood Cliffs, NJ: Prentice Hall.

IEEE Transactions on Communications. Retrieved from May 6, 2003, http://www.comsoc.org/pubs/jrnal/transcom.html

IEEE Information Theory Society Homepage. Retrieved from May 6, 2003, http://www.itsoc.org

IEEE Signal Processing Society Homepage. Retrieved from May 6, 2003, http://www.eee.org/organizations/society/sp

Digital Divide

Jaime J. Dávila, *Hampshire College*

INTRODUCTION

Broadly speaking, the term "digital divide" refers to the difference in access to computers and the Internet that exists among different groups. In the following paper I will present a historical overview of this issue, its most important characteristics, and how it manifests domestically and internationally. In addition, I will discuss several ways in which the situation is being dealt with and propose possible solutions.

On October 10, 1996, President Clinton and Vice President Gore addressed the public at the Knoxville Auditorium Coliseum (Clinton & Gore, 1996). The main topic of their presentation dealt with the government's effort to build a bridge to the 21st century. During the conversation, they talked about the existing digital divide, demonstrating the importance being given to this issue by high-ranking government officials. Since then, the term has come to be synonymous with the different levels of access to the Internet experienced by different groups of people. In the first of several government policy documents dealing with this issue, the importance of closing the digital divide is defined as important because "... individuals' economic and social well-being increasingly depends on their ability to access, accumulate, and assimilate information" (United States Department of Commerce, 1995).

In the following pages I will present a more formal definition for the term digital divide, as well as how it manifests among different groups, both in the United States and internationally. Finally, I will present a series of conclusions and recommendations regarding possible ways to solve this important issue.

GOVERNMENT DEFINITIONS

The first official report by the federal government having to do with inequities in access to the Internet was made available to the public in July 1995 (United States Department of Commerce, 1995). The NTIA (National Telecommunications and Information Administration), an agency within the United States Department of Commerce, was created in 1978 to serve as the President's principal adviser in matters related to telecommunications and information policy issues. In the first of a series of reports, the NTIA reported that specific groups across the nation were lagging behind in their home access to computers and connectivity to the Internet. The most important variables defining these groups were economic status, ethnicity, age, educational level, and geography. Three years later, the Department of Commerce (1998) reported that although the number of U.S. houses with computers and connections to the Internet had increased, the percentage difference, among groups based on ethnicity, income, education level, and age, between those with and those without Internet access had, in fact, grown. In addition, they reported that single-parent households were also considerably less likely to have Internet connectivity.

The last of the reports published under the *Falling Through the Net* series was published in 1999. *Falling Through the Net: Defining the Digital Divide* (Department of Commerce, 1999b) reported that, although 40% of households had computers and 25% of households had Internet connectivity, the differences between the different groups continued to exist. According to the report,

> The 1998 data reveal significant disparities, including the following: Urban households with incomes of $75,000 and higher are more than twenty times more likely to have access to the Internet than rural households at the lowest income levels, and more than nine times as likely to have a computer at home. Whites are more likely to have access to the Internet from home than Blacks or Hispanics have from any location. Black and Hispanic households are approximately one-third as likely to have home Internet access as households of Asian/Pacific Islander descent, and roughly two-fifths as likely as White households. Regardless of income level, Americans living in rural areas are lagging behind in Internet access. Indeed, at the lowest income levels, those in urban areas are more than twice as likely to have Internet access than those earning the same income in rural areas.

In fact, for many of these groups, the percentage difference between having and not having Internet access was found to have increased through the years.

> The gaps between White and Hispanic households, and between White and Black households, are now approximately five percentage points larger than they were in 1997. . . . The digital divides based on education and income level have also increased in the last year alone. Between 1997 and 1998, the divide between those at the highest and lowest education levels increased 25 percent, and the divide between those at the highest and lowest income levels grew 29 percent. (Department of Commerce, 1999b)

On a more positive note, the report found that the differences based on ethnicity disappeared for annual household incomes above $75,000. This suggested that lower equipment costs could make the divide disappear from other economic sectors.

In February 2002, under a new federal government, the NTIA generated a report on the issue of the digital divide with a very different outlook. Based on the most recent census data, a record number of Americans were making use of computers and the Internet. In a new report, the United States Department of Commerce (2002) reported

> Between December 1998 and September 2001, Internet use by individuals in the lowest-income households (those earning less than $15,000 per year) increased at a 25 percent annual growth rate. Internet use among individuals in the highest-income households (those earning $75,000 per year or more) increased from a higher base but at a much slower 11 percent annual growth rate . . . Between August 2000 and September 2001, Internet use among Blacks and Hispanics increased at annual rates of 33 and 30 percent, respectively. Whites and Asian American and Pacific Islanders experienced annual growth rates of approximately 20 percent during these same periods. . . . Over the 1998 to 2001 period, growth in Internet use among people living in rural households has been at an average annual rate of 24 percent, and the percentage of Internet users in rural areas (53 percent) is now almost even with the national average (54 percent).

CLASS VERSUS ETHNICITY

Looking at the possible exclusion of a group of citizens from benefiting from technology shows how important it is to identify factors influencing this exclusion. Two of the most commonly cited factors are class and ethnicity. Novak and Hoffman (1998) found that income is a better predictor of access to computers at home, and that educational level is a better predictor of having access to computers at work. At the same time, the Benton Foundation (1998) reported that for households earning under than $40,000/year, white Americans are close to twice as likely to have computers or Internet access at home than African Americans. The numbers are similar when comparing white Americans with Latinos (Becht, Taglang, & Wilhelm, 1999) : For households earning under than $15,000, the gap between non-Hispanic White and Hispanic households rose substantially, from 5.6% to 8.1% between 1994 and 1998. For households earning between $15,000 and $34,999, the disparities between non-Hispanic White and Hispanic households increased by 46% or (4.0 percentage points). A White, two-parent household earning under than $35,000/year is nearly four times more likely to have Internet access than Hispanic households in that same income category.

One proposed solution to this problem is to effectively increase the use of more publicly available Internet access points. For example, those citizens who lack Internet connectivity at home could gain it at schools, libraries, or community centers. Although in theory this seems to be a good idea, it is one that has to be approached, analyzed, and improved on carefully. Although 98% of schools in the United States have computers, the ratio of students per computer is 30% higher in schools where 80% or more of the student body participates in free or reduced-cost lunch programs. In schools where more that 90% of the student body is from minority groups, the student-to-computer ratio is 74% higher (Educational Testing Service, 1997). The problem is even bigger when measuring Internet access. Among schools with a student body composed of 50% or more minority students, Internet connectivity is available in 22% fewer schools (National Center for Education Statistics, 1997). A study carried out by the Educational Testing Service also found that schools in low-income areas are close to 25% less likely to have Internet connectivity than those in high-income areas (Benton Foundation, 1998).

With regard to public libraries, a study performed by the U.S. National Commission on Libraries and Information Science (1998) revealed that although a high percent of public libraries across poverty classifications have public Internet access (92.8%–95.6%), the percentage of libraries located in poor communities is very low (2.0%). In addition, although schools, public libraries, and community centers can provide valuable connectivity services, access to the Internet at public spaces will necessarily have characteristics different from access at home. Issues based around waiting in queues, time limits, and privacy can all lead to less than positive experiences for those who do not have Internet access from home. For example, even though most unemployed people access the Internet from outside their homes, most unemployed people using the Internet to look for jobs were doing it from their houses (U.S. Department of Commerce, 2000).

OTHER DISTINGUISHING FACTORS IN THE UNITED STATES

There is a notable difference in computer ownership and use of the Internet depending on geographic location. In 2001, although 57.4% of households in urban houses outside of inner cities had access to the Internet, the number was 3.5 percentage points lower for rural households, and 7.7 percentage points lower for inner city households (United States Department of Commerce, 2002). This number is particularly significant because it includes access to the Internet at any location, inside or outside the home, and translates to less than half of households in inner cities having access to the Internet.

In terms of gender, computer usage and Internet access has come to match general population percentages in the United States (United States Department of Commerce, 2002). In 2001, 53.9% of men and 53.8% of women had access to the Internet. There are other ways, though, in which women are being left behind in terms of technology. One of them, regarding women connected to the Internet in countries other than the United States, will be discussed in this chapter in a later section, on the international digital divide. In addition, there are significant differences in number of women versus men in technology education.

The American Association of University Women Educational Foundation (2000) reported that although they comprised 51% of the U.S. population, in the year 2000, females obtained only 28% of the degrees in computer science (down from 34% in 1984) and comprised only 20% of the IT workforce.

Finally, serious consideration should be given to the way in which disabled people access the Internet. Those who are visually impaired, or have problems working with a mouse as a pointing device, can have great difficulty navigating and taking advantage of the opportunities offered by technology as a communication medium. In 2000, the U.S. Department of Commerce (2000) reported that although 56.7% of U.S. citizens accessed the Internet (either from inside or outside their homes), only 28.4% of people with disabilities did. Only 22.5% of people with difficulty using their hands and 21.1% of people with vision problems used the Internet. In September 2000, President Clinton assigned $16 million to the Department of Education to promote technology programs for the disabled, and an additional $9 million to Americorps volunteers working with Internet for the disabled. In addition, modifications made to the U.S. Rehabilitation Act now require government web pages to meet certain criteria to make them more accessible to people with disabilities (CNN, 2000a).

THE TECHNOLOGY S-CURVE PROCESS

Recent analysis of the digital divide problem argues that the differences seen among different groups are closing and will continue to close as more and more people come online. This can partly be explained by standard processes of technology diffusion. Rogers (1983) defined five types of innovation adopters: innovators, early adopters, early majority, late majority, and laggards. He further argued that all technologies have a diffusion rate into general society. At first very few people use the technology. With time, more individuals start to use the technology, until a saturation point is reached. After this saturation point, the rate at which additional people start using the technology for the first time decreases, with the total number of users staying close to constant. According to Rogers, the rate at which a new technology is adopted by the general population is a factor of five characteristics: its relative advantage over the technologies that it replaces; how compatible it is with the values and experiences of the adopters; how complex the technology is perceived to be; how easy it is for people to try out; and how easy it is for others to see the results of adopting it. Many of these factors were, in fact, empirically identified by the general population in a survey carried out by Katz and Aspden (1997). Given these factors, only a small percentage of the population is bound to try any new technology soon after it is introduced. But as prices drop and the effects of the technology are made public, more people start using it. Figure 1, reproduced with small modifications from the United States Department of Commerce (2000), shows a typical S-curve. If this pattern applies to computers and the Internet, then the current digital divide will naturally close with time. Although Leigh and Atkinson (2001) report that other technologies introduced during the 20th century (such as telephones, radios, and

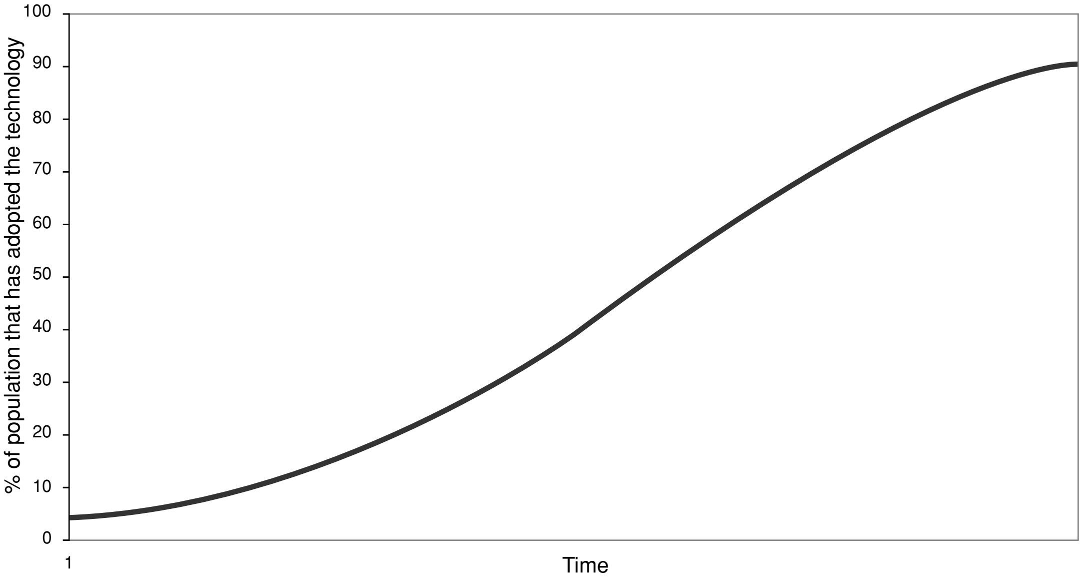


Figure 1: Typical technology S-curve.

television) grew to almost 100% penetration among U.S. households, the trend might not apply to Internet connectivity. As evidence, Internet penetration reached 30% of the U.S. population in seven years, making it a faster growing technology than telephones, television, or video cassette recorders (VCR; Leigh & Atkinson, 2001), but it failed to have any significant increase between November 2000 and April 2002 (Taylor, 2002). This can mean one of two things: Either we have reached the top of the S-curve, in which case the percentage of users will not continue to expand, or Internet penetration will behave in differently than the S-curves seen with other technologies, in which case the S-curve theory of penetration cannot be used to make predictions about the future expansion of the Internet.

ACCESS VERSUS SKILLS

In addition to access to computers and the Internet, the digital divide includes issues around the training and skills necessary to take advantage of both of these resources once they are attained. Warschauer (2002), for example, has documented the results of several Internet connectivity projects. In one of them, four towns in Ireland were given government funds to use to turn themselves into "Information age towns." With less money, three of these towns were able to create more successful programs than the town that received the highest per capita amount. The distinguishing factor was "... developing awareness, planning and implementing effective training, and setting up processes for sustainable change, rather than merely on purchase of equipment" (Warschauer, 2002).

Warschauer discusses a similar example in a classroom setting in Hawaii, where two schools made clearly different use of computers and technology (Warschauer, 2000). Although both schools used technology to improve the educational opportunities available to their students, differences were evident. The school in the more economically affluent community used computers to develop academic and scientific skills, where the students in the poorer school were involved with tasks more directed toward the workforce. This difference was matched by a series of goals and expectations set by the teachers at each of the schools.

Being able to efficiently use the Internet, with its multitude of information, requires a series of different but related skills. Carvin (2000) identifies six types of literacy skills needed by a user in order to take full advantage of the Internet: basic literacy (the ability to read and write); functional literacy (the ability to apply basic literacy to everyday tasks); occupational literacy (the skills necessary to succeed in a professional setting); technological literacy (the ability to use technological tools); information literacy (the ability to determine the quality of informational sources); and adaptive literacy (the ability to develop new skills).

Parallel to the issue of skills development, the relevance of available content can control the benefit any community can obtain from being connected to the Internet. This becomes particularly important for low-income communities, for whom this new medium might be one of very few avenues for improvement. An audit performed by the Children's Partnership (Lazarus & Mora, 2000) revealed that technologically underserved U.S. Citizens found the Internet lacking in local information and cultural diversity. In addition, their study found that although 22% of the population of the United States does not have proper "everyday" literary skills, most of the text available online is written assuming that the audience has at least an average level of literacy.

THE INTERNATIONAL DIGITAL DIVIDE

As is to be expected, the number of people connected to the Internet outside of the United States varies greatly, from numbers as high as 60%, for the population in Canada (Mandel, 2002), which is higher than in the United States, to less than 0.04% in Ethiopia (International Telecommunications Union Statistics, 2002). Although many countries in Europe, Asia, and the Pacific are starting to catch up with the United States and Canada in terms of Internet use, they are still lagging (Campbell, 2001). The problem is even greater when we look at countries in Africa and Latin America.

Norris (2001) analyzed data on Internet usage for Portugal, Greece, Germany, Spain, France, Belgium, Austria, Ireland, Italy, the United Kingdom, Luxembourg, Holland, Finland, Denmark, and Sweden. In all these countries, income, educational level, gender, age, and type of occupation (managerial versus manual work) was a strong indicator of Internet connectivity. The problem of a gender-based digital divide, now apparently absent in the United States, is a major problem in countries outside Europe. Of the total number of people connected to the Internet, only 22%, 38%, and 6% are women from the regions of Asia, Latin America, and the Middle East, respectively (Hafkin & Taggart, 2001).

There are also drastic domestic differences within developing countries. Income, education, geographic location, and gender all seem to influence connectivity into such required communication links as telephones (Grace, Kenny, & Qiang, 2001). Most organizations evaluating a nation's readiness to deploy and take advantage of the Internet identify five critical areas: the availability and cost of connection lines, such as telephones and cellular phones; government leadership in areas related to the Internet; current computer infrastructure, such as number of computers per capita; training of the local workforce; and local climate for electronic commerce.

Connectivity

Many of the differences in access to the Internet run parallel to access to telephone lines. In the United States, where close to 54% of the citizens have access to the Internet, 94% of households have telephone connections (U.S. Department of Commerce, 1999b). In contrast, Africa, which lags any other continent in Internet usage, has a total of 14 million telephone lines (fewer telephone lines than either Manhattan or Tokyo (Nkrumah, 2002)). Ethiopia, which has only 1.1 Internet users per 10,000 citizens, has only 0.32 telephone lines per 100 habitants (BBC News, 2000). This problem is particularly severe in light of the difficulty of providing connectivity to remote parts of a country. This creates a domestic digital divide in countries where towns and villages are far removed from urban centers.

Another factor related with telephone lines is the cost per minute of connecting to the Internet. In many countries, charges for telephone calls are accrued on a per-minute basis. The Organization for Economic Cooperation and Development (2000) has found strong correlations between the price per minute for telephone calls and the number of Internet hosts in different countries. This is particularly important in light of another of the factors that seems to foster domestic connectivity: local climate for electronic commerce.

Government Leadership

Different national governments have implemented a variety of plans to reduce their domestic digital divide. The one common thread among all of them is their signaling that access to the Internet is an important national issue. In the United States, President Clinton (White House Web Site, 2000) presented a budget in February 2000 that included several key items designed to increase Internet access across the nation. Among other things, his proposal called for $2 billion in tax incentives for private companies donating hardware and training to technology centers; $150 million to train new teachers in the best ways to use technology; $100 million to create community technology centers; $25 million to encourage the private installation of Internet connectivity to underconnected communities; and $10 million to train native Americans for careers in information technologies. In Brazil, the government intends to provide Internet terminals at post offices in cities with 6,000,000 or more habitants (Rebeto, 2001a). In addition, it is providing Internet-ready home computers for as little as $15/month (Bebeto, 2001b). The Australian government has put in place a series of initiatives to bring both connectivity and training to all of its population, at a total cost of close to $426 million (Australian National Office for the Information Economy, 2000).

Current Computer Infrastructure

In many countries, we see a close correlation between the number of computers per person and the number of people accessing the Internet. On one end of the spectrum is the United States, where in the year 2001, 66% of the population owned computers and 54% of the population had access to the Internet (U.S. Department of Commerce, 2002). At the other end of the spectrum, as reported by Tran and Barrett (2001), in Bolivia only 0.75% of the population owns a computer and only 0.98% have some type of access to the Internet (NUA, 2002). One simple explanation for this relationship is that a high number of people access the Internet from their households. Additionally, local companies are bound to facilitate Internet points of access if they feel there is an audience ready to subscribe to that type of service.

Technology Training

It is not enough to have a computer and network connectivity to take advantage of the Internet. Potential users need to understand what they have to gain from using the medium, and they need to know how to access its features. Both national governments and private institutions can play important parts in delivering this knowledge to citizens. In Jordan, for example, the government has established a program to teach Information Technology at academic institutions across the country (Niewiaroski, 2002). In addition, they have started partnerships with private companies, such as AOL, Microsoft, Oracle, and Cisco, to retrain computer graduates and keep them up to date with the latest technologies. In Mexico, the Technological Institute of Monterrey now offers technology courses on the latest IT topics to its own population as well as to other Latin American countries (McConnel International, 2001). Many of these initiatives share the common thread of being partnerships between public, private, and educational institutions.

E-commerce Climate

A final characteristic for successful national Internet inclusion is the ease with which companies can establish electronic commerce and transactions. This benefits nations in three ways: It provides additional avenues for job creation and national economic development (Asian Productivity Organization Secretary-General, 2000); it stimulates local and international entrepreneurs to assist with other items necessary for generalized Internet participation, because they gain from an increase in online customers; and it provides incentives for citizens to acquire training in information technologies, because companies will now require a trained workforce.

At the same time, nations that have a stable base of Internet users, as well as a developed computer infrastructure, can more readily attract e-business companies. Reports by the Organization for Economic Cooperation and Development (2000) demonstrate strong correlations between the number of Internet users, the cost of connecting, and the number of secure servers for its member nations.

In this way, then, connectivity, government leadership, computer infrastructure, a trained local workforce, and a positive electronic commerce climate form the cornerstones of Internet development for any country. All these factors feed off each other in a systemic process. For example, local governments have the primary responsibility for providing connectivity, but they usually come into partnerships with private industries in order to facilitate the rigorous task of wiring a nation. Private industries, at the same time, take benefit of government subsidies in exchange for providing hardware and training to the general population.

ADVANTAGES OF BEING CONNECTED

The advantages provided by the Internet are many, but all deal with issues of acquiring and communicating information. The most immediate of these advantages vary depending on the user and his or her location. In the United States, for example, 53.9% of unemployed people who use the Internet from their homes and 29.8% of unemployed people who use the Internet from outside their homes use it to look for new jobs (U.S. Department of Commerce, 1999a). Additionally, between 36 and 38% of people who use the Internet from any location are either taking online courses or doing academic research.

For countries around the world, taking advantage of new technological opportunities can be crucial for national economic development. Campbell (2001) has found a strong correlation between the number of technology users and the percentage of growth in employment.

Citizens of a connected nation can follow the work of their elected representatives more easily. Many elected government officials can be contacted via the Internet. The U.S. House of Representatives can be reached at http://www.house.gov. The U.S. Senate can be reached at http://www.senate.gov. The State of Texas Legislature can be reached at http://www.capitol.state.tx.us/fyi/fyi.htm. These Web sites not only provide information, they also allow for easier contacting of elected officials. A new way for citizens to participate in government is evolving out of these communication media. Information about government rules, regulations, and benefits can be found online. For example, the state of Oregon provides such information at http://www.oregon.gov. This type of Web site directs citizens to the correct office for the services they might be looking for and, in some cases, allows visitors to request services online. Labor unions now use the Internet to inform and organize their members (CNN, 2000b). Union members are thus kept better informed and are in closer contact with the people they elect to act as their representatives.

In October 2000, 45% of the African-Americans who used the Internet reported looking for health information (Associated Press, 2000). High school technology education programs are training young programmers living in urban neighborhoods, allowing them to enter the workforce and command salaries of up to $150/hr (Clewley, 2001). This offers a group of typically disenfranchised citizens skills that are in high demand, turning technology into an engine capable of propelling people into better lifestyles. In Sierra Leone, a local entrepreneur has started a program where young rebel soldiers are able to trade their arms for computer training, with the goal of turning them into the center of a software outsourcing business within 3 years (Hermida, 2002a).

In Argentina, as part of a series of activities designed to put pressure on government officials, protesters have set up a number of Web sites documenting their dissatisfaction with recent economic rulings (d'Empaire, 2002). In Kosovo, Albanians and Serbians are coming together and using the Internet to make their environmental problems known (Hermida, 2002a). By using the Internet as a medium to communicate their problems to others, these groups and individuals are now better able to obtain assistance for their causes.

The number of programs like those mentioned above is too extensive to list here. These programs, and others like them, have very different flavors and approaches, depending on the countries in which they are being implemented. What all of them have in common is their use of the Internet as a powerful communication mechanism, as well as the door to new economic opportunities. As such they can serve as tools for increasing equality, bringing new opportunities to disadvantaged communities. If, on the other hand, the divide between those who have access to the Internet and those who do not is not closed, it may well serve as a mechanism to exacerbate economic and social differences between the haves and the have nots.

SOLUTIONS TO THE DIGITAL DIVIDE

Given the facts outlined throughout this chapter, a pattern of solutions to the problem of a digital divide starts to emerge. In most nations, and in particular in the United States, the government must be an active participant in efforts to provide equal access to members of different communities. As an example of a similar process, telephone penetration in the United States increased dramatically during the 20th century. An important reason for this phenomenon was federal government legislation, such as the Communications Act of 1934, which defined the ways in which private companies were to benefit from their efforts to provide telephone connectivity to the United States. By participating in the creation of the necessary infrastructure, private companies gain the creation of potential customers.

Federal governments can also help to close the digital divide by directly allocating funds to entities wiring and equipping computer centers in public spaces. An example of such a program is the Neighborhood Networks program of the U.S. Department of Housing. Since 1995, this program has been providing funds for the establishment of computer centers within low- or moderate-income multifamily housing. A similar initiative, commonly known as the E-rate program, allocated federal funds to the development of computer access at public and private schools and libraries. A study performed by the Urban Institute (Puma, Chaplin, & Pape, 2000) reported that most of these funds went into public schools, particularly in the poorest communities.

In the international arena, governments can use special incentives and regulations to create market environments favorable to private institutions. These institutions are then stimulated to help develop the physical and human infrastructure that allows citizens to become connected. As the governmental budget of a particular nation decreases, the economic contribution of the private sector becomes more important.

For those individuals and groups whose economic situation does not allow them to gain access to the Internet at home, libraries and community centers will become critical points of Internet connectivity. Once again, public and private coalitions are central to continuing to implement these possibilities. Government entities can provide incentives to the private sector to donate equipment and/or training to public schools, libraries, and community centers. Programs such as CTCNet and Americorps have invested in both creating community centers and training the community how to take advantage of technology. This type of initiative needs to receive both public and private funds to continue doing the job. In both the domestic and international arena, the success of these programs will be dependent on integrating these efforts with the search and implementation of solutions to the problems of each community. Once in progress, these types of government/community programs have the ability to empower citizens who have traditionally been marginalized. A study on the effect of the CTCNet program (Chow, Ellis, Mark, & Wise, 1998) found that females, people of all ages, and members of minority groups all made extensive use of community computer centers associated with this program. Participants used the Internet for such things as expanding their skills and looking for jobs. In addition, interviewees reported an increase in personal effectiveness, as well as a sense of community building. Of the people polled in the survey, 94% reported positive feelings about using the community centers.

Partnerships between public and private entities have demonstrated great ability to assist in connecting individuals in low-income communities. The *Computers for Youth* program (http://www.cfy.org), for example, has created a partnership between private companies, branches of the federal government, universities, and other educational institutions in order to place computers in predominantly Black and Hispanic households.

Governments and institutions in developed countries have the ability to greatly assist developing countries eliminate internal and international digital divides. The Global Digital Opportunity Initiative is a result of studies prepared by the Group of Eight Nations (G8). In it, the Markle Foundation, for example, has come into partnership with the United Nations Development Program and several key computer technology companies to provide pro bono consulting services to countries interested in creating strategies for local Internet development (Sullivan, 2002).

Colleges and universities are also important players in providing Internet access to those who currently lack it. Their ability to both establish community programs and do research about the subject makes them an important information clearinghouse. Of particular interest is their ability to embed training into community activities, and also their ability to disseminate results and findings. An example of this is currently being carried out by the Stanton/Heiskell Center for Public Policy in Telecommunications and Information Systems, at the City University of New York Graduate School. In 1990, this research center placed networked computers in the homes of 125 students from typically underserved communities (Stanton/Heiskell Center, 1998). At the conclusion of their 7-year program, a series of findings was disseminated to the public. In addition, newer pilot programs were established to examine questions that arose from their original research. Not surprisingly, public and private institutions were instrumental in funding and facilitating these programs.

In relation to access for disabled people, economic stimulus and publicity should be used to encourage companies to develop accessible software and hardware. An example of increased software accessibility will be seen in additions to Macromedia's Flash, which is widely used to create interactive Web sites. Jason Smith, an independent software programmer, has created an addition to the standard Flash program that will allow computer screen-reading packages to read captions in the Flash animations (Delgado, 2002).

Finally, a certain sector of the population will adopt the use of Internet connectivity only after being educated about its advantages. Some people, particularly in the United States, have reported not being connected because they do not see the utility of using the Internet (U.S. Department of Commerce, 2000). Given the number of uses the Internet can have, it is safe to say that these people are not aware of the opportunities the medium offers them. Schools, community centers, libraries, and other public spaces can help with this population by holding programs that communicate the benefits of the Internet to those who have the infrastructural requirements to enjoy it but are not convinced of its utility. These public groups, in turn, will need public and private economic assistance to carry out this type of educational program.

CONCLUSION

The digital divide is a problem affecting broad sections of our communities, both in the United States and internationally. Because of the Internet's potential to benefit its users, not solving this divide can cause the difference in quality of life to increase dramatically through the years. Efforts to solve the issue of the digital divide must deal with developing infrastructures, providing training, and empowering communities to integrate the Internet into their normal everyday activities. The results of projects carried out at different levels across the world indicate that programs aimed at closing the digital divide are most effective when they integrate the participation of governments, private institutions, community groups, and the general citizenry.

GLOSSARY

Connectivity Refers to the availability of physical media through which communication can take place.

Domestic/national Unless otherwise specified, domestic and national terms are relative to the United States of America.

Ethnicity Affiliation based on membership in a particular racial, linguistic, or cultural group.

Infrastructure The resources necessary to for participation in a given activity.

CROSS REFERENCES

See *Developing Nations; Digital Economy; Global Issues; Legal, Social and Ethical Issues; Politics.*

REFERENCES

American Association of University Women Educational Foundation (2000). *Tech-savvy: Educating girls in the new computer age*. Iowa City, IA: American Association of University Women Educational Foundation.

Asian Productivity Organization Secretary-General (2000). *International conference on productivity in the e-age, inaugural address*. Retrieved July 22, 2002, from http://www.apo-tokyo.org/sgstatem/0a_sg_20001122.htm

Associated Press (2000). *Blacks go to Net for information*. Retrieved July 22, 2002, from http://www.wired.com/news/business/0,1367,39614,00.html

Australian National Office for the Information Economy (2000). *Digital divide*. Retrieved July 22, 2002, from http://www.noie.gov.au/projects/access/community/digitaldivide/Digitaldivide.htm

BBC News (2000). *When the Web is not world-wide*. Retrieved July 22, 2002, from http://news.bbc.co.uk/hi/english/sci/tech/newsid_843000/843160.stm

Becht, D., Taglang, K., & Wilhelm, A. (1999). *The digital divide and the US hispanic population*. Retrieved July 22, 2002, from http://www.benton.org/DigitalBeat/db080699.html

Benton Foundation (1998). *Losing ground bit by bit*. Retrieved July 22, 2002, from http://www.benton.org/Library/Low-Income/

Broberg, B. (2001). *Provail helps disabled bridge the digital divide*. Retrieved July 22, 2002, from http://seattle.bizjournals.com/seattle/stories/2001/08/20/focus1.html

Campbell, D. (2001). Can the digital divide be contained? *International Labour Review*, *1*, 119–141.

Carvin, A. (2000). *Literacy and the digital divide*. Retrieved July 27, 2002, from http://www.educause.edu/pub/er/erm00/articles006/erm0063.pdf

Chow, C., Ellis, J., Mark, J., & Wise, B. (1998). *Findings from a national survey of users of community technology centers*. Retrieved April 28, 2002, from http://www.ctcnet.org/impact98.htm

Clewley, R. (2001). *Programming a way out of poverty*. Retrieved July 22, 2002, from http://www.wired.com/news/school/0,1383,45922,00.html

Clinton, W., & Gore, A. (1996). *Remarks by the President and the Vice President to the people of Knoxville*. Retrieved July 22, 2002, from http://www.ntia.doc.gov/ntiahome/101096clinton.htm”http://www.ntia.doc.gov/ntiahome/101096clinton.htm

CNN.com (2000a). *Clinton pushes to help disabled bridge ‘digital divide’*. Retrieved July 22, 2002, from http://www.cnn.com/2000/ALLPOLITICS/stories/09/21/clinton.digital

CNN.com (2000b). *Labor unions turn to cyberspace to organize, inform*. Retrieved July 22, 2002, from http://www.cnn.com/2000/TECH/computing/09/04/unions.internet.ap/index.html

Delgado, L. (2002). *Flash news flash: It’s accessible*. Retrieved July 22, 2002, from http://www.wired.com/news/culture/0,1284,51638,00.html

D’Empaire, A. (2002). *Banging saucepans on the Net*. Retrieved July 22, 2002, from http://www.wired.com/news/politics/0,1283,50920,00.html

Educational Testing Service (1997). *Computers and classrooms: The status of technology in U.S. schools*. Princeton, NJ: Educational Testing Service.

Grace, J., Kenny, C., & Qiang, C. (2001). *Information and communication technologies and broad-based development: A partial review of the evidence*. Retrieved July 22, 2002, from http://wbln0018.worldbank.org/ict/resources.nsf/c701be8cd977a3cc852569490049e219/1dda15ac468eb637852569fb0053cb77/$FILE/graceetal0214.doc

Hafkin, N., & Taggart, N. (2001). *Gender, information technology, and developing countries: An analytical study.* Retrieved July 22, 2002, from http://www.usaid.gov/wid/pubs/hafnoph.pdf

Hermida, A. (2002a). *Child soldiers to swap guns for PCs*. Retrieved July 22, 2002, from http://news.bbc.co.uk/hi/english/sci/tech/newsid_1886000/1886248.stm

Hermida, A. (2002b). *Internet unites Kosovo foes*. Retrieved July 22, 2002, from http://news.bbc.co.uk/hi/english/sci/tech/newsid_1939000/1939121.stm

International Telecommunications Union (2000). *International Telecommunications Union statistics.* Retrieved July 22, 2002, from http://www.itu.int/ITU-D/ict/statistics/index.html

Katz, J., & Aspden, P. (1997). Motivations for and barriers to Internet usage: Results of a national public opinion survey. In *Internet Research: Electronic Networking Applications and Policy* (Vol. 7, pp. 170–188). Bradford, England, U.K.: MCB University Press.

Lazarus, W., & Mora, F. (2000). *On-line content for low-income and under served Americans: The digital divide’s new frontier*. Retrieved July 23, 2002, from http://www.childrenspartnership.org/pub/low_income/index.html

Leigh, A., & Atkinson, R. (2001). *Clear thinking on the digital divide*. Retrieved July 22, 2002, from http://www.ndol.org/documents/digital_divide.pdf

Mandel, C. (2002). *Connectivity kings: Oh, Canada*. Retrieved July 22, 2002, from http://www.wired.com/news/business/0,1367,51678,00.html

McConnel International (2001). *Ready, Net, go! Partnerships leading the global economy*. Retrieved July 22, 2002, from http://www.mcconnellinternational.com/ereadiness/ereadinessreport2.htm

National Center for Educational Statistics (1997). *Advanced Telecommunications in U.S. Public Elementary and Secondary Schools, NCES 97–944*. Washington, DC: National Center for Educational Statistics.

Niewiaroski, D. (2002). *Jordan’s new model IT economy*. Retrieved, from http://www.arabdatanet.com/news/DocResults.asp?DocId=2882

Nkrumah, G. (2002). *Digital divide*. Retrieved July 22, 2002, from http://www.ahram.org.eg/weekly/2000/492/in3.htm

Norris, P. (2001). *Digital divide: Civic engagement, information poverty, and the Internet worldwide.* Cambridge, UK: Cambridge University Press.

Novak, P., & Hoffman, D. (1998). *Bridging the digital divide: The impact of race on computer access and Internet use*. Retrieved July 22, 2002, from http://elab.vanderbilt.edu/research/papers/html/manuscripts/race/science.html

NUA (2002). *How many online?* Retrieved July 22, 2002, from http://www.nua.ie/surveys/how_many_online/index.html

Organization for Economic Cooperation and Development (2000). *Working party on telecommunications and information services policies, pricing and e-commerce*. Retrieved July 22, 2002, from http://www.olis.oecd.org/olis/2000doc.nsf/LinkTo/DSTI-ICCP-TISP(2000)1-FINAL

Puma, M., Chaplin, D., & Pape, A. (2000). *E-rate and the digital divide: A preliminary analysis from the integrated studies of educational technology*. Retrieved July 22, 2002, from http://www.urban.org/education/erate.html

Rebeto, P. (2001a). *Casting a wider net in Brazil*. Retrieved July 22, 2002, from http://www.wired.com/news/politics/0,1283,45526,00.html

Rebeto, P. (2001b). *Brazil counting on a net gain*. Retrieved July 22, 2002, from http://www.wired.com/news/culture/0,1284,41785,00.html

Rogers, E. (1983). *Diffusion of innovations* (3rd ed.). New York: The Free Press.

Stanton/Heiskell Center for Public Policy in Telecommunications and Information Systems (1998). *Project TELL: Telecommunications for learning*. Retrieved July 22, 2002, from http://web.gc.cuny.edu/shc/docu2.htm

Stoecker, R., & Stuber, A. (1997). Limited access: The information superhighway and Ohio's neighborhood-based organization—The Urban University and Neighborhood Network. *Computers in Human Services, 14,* 39–57.

Sullivan, B. (2002). *UN initiative to help developing countries boost IT infrastructure*. Retrieved July 22, 2002, from http://www.computerworld.com/careertopics/careers/story/0,10801,68042,00.html

Taylor, H. (2002). *Internet penetration at 66% of adults (137 million) nationwide*. Retrieved July 22, 2002, from http://www.harrisinteractive.com/harris_poll/index.asp?PID = 295

Tran, Q., & Barrett, E. (2001). *Access in Bolivia. Trends in Latin American networking*. Retrieved July 22, 2002, from http://lanic.utexas.edu/project/tilan/reports/rtf359/bolivia1.html

United States Department of Commerce (1995). *Falling through the net: A survey of the "have nots" in rural and urban America*. Retrieved July 22, 2002, from http://www.ntia.doc.gov/ntiahome/fallingthru.html

United States Department of Commerce (1998). *Falling through the net II: New data on the digital divide*. Retrieved July 22, 2002, from http://www.ntia.doc.gov/ntiahome/net2/falling.html

United States Department of Commerce (1999a). *Fact sheet: Americans using Internet for many tasks*. Retrieved July 22, 2002, from http://www.ntia.doc.gov/ntiahome/digitaldivide/factsheets/usage.htm

United States Department of Commerce (1999b). *Falling through the net: Defining the digital divide.* Retrieved July 22, 2002, from http://www.ntia.doc.gov/ntiahome/fttn99/contents.html

United States Department of Commerce (2000). *Falling through the net: Toward digital inclusion. Index of charts and tables*. Retrieved July 22, 2002, from http://www.ntia.doc.gov/ntiahome/fttn00/chartscontents.html

United States Department of Commerce (2002). *A nation online: How Americans are expanding their use of the Internet*. Retrieved July 22, 2002, from http://www.ntia.doc.gov/ntiahome/dn/index.html"http://www.ntia.doc.gov/ntiahome/dn/index.html

U.S. National Commission on Libraries and Information Science (2000). *Public library Internet study*. Retrieved July 22, 2002, from http://www.nclis.gov/statsurv/2000plo.pdf

Warschauer, M. (2000). *Technology and school reform: A view from both sides of the track*. Retrieved July 22, 2002, from http://epaa.asu.edu/epaa/v8n4.html

Warschauer, M. (2002). *Reconceptualizing the digital divide*. Retrieved July 23 2002, from http://www.firstmonday.dk/issues/issue7_7/warschauer/

White House Web Site (2000). *The Clinton–Gore administration: From digital divide to digital opportunity*. Retrieved July 23, 2002, from http://clinton4.nara.gov/WH/New/digitaldivide/digital1.html

Digital Economy

Nirvikar Singh, *University of California, Santa Cruz*

INTRODUCTION

> IT and the Internet amplify brain power in the same way that the technologies of the industrial revolution amplified muscle power.
>
> —Bradford DeLong, Professor of Economics, University of California, Berkeley, quoted in Woodall (2000, p. 6)

The purpose of this chapter is to explain what the digital economy is and how it fits into broader economic trends that are shaping the economies of the United States and other countries. Essentially, the digital economy refers to the use and impact of digital information technology in various forms of economic activity. As such, the term includes more specific activities such as e-commerce and e-business and is closely related to terms such as knowledge economy and information economy.

All groups in the economy are affected by the pervasive use of information technology; this includes consumers, business firms, and governments. Activities that are not directly commercial, such as personal communications, are also affected. The fundamental driving force is the falling costs of processing, storing, and transmitting information that has been put into digital electronic form. These declines in costs have made innovations possible that permit easy and widespread communication over extensive networks, the existence of large and rich databases of information and knowledge that are freely or easily accessible, and the ability to conduct most or all stages of economic transactions over long distances, without relying on alternative methods of information exchange.

The further results of these expanded capabilities for accessing, using, and sharing information in digital form include new and more efficient ways of organizing markets; new and more efficient methods for businesses to communicate and transact with each other and with consumers, employees, and job seekers; dramatically lower costs for individuals in locating or gathering information of all kinds, including market- and product-related information; changes in the organization of business firms and in their strategic behavior; and changes in the overall societal organization of work, leisure, and general communities of shared interest.

The essay is organized as follows. We provide some data on the size and growth of the digital economy and e-commerce and discuss the measured impacts of information technology on the economy as a whole, as well as some of the problems of measurement. We describe changes in the nature, structure, and performance of firms and markets as digital technologies help make information more ubiquitous and as information increases in importance as an economic good. We examine several aspects of government policy with respect to the digital economy, including contracting, privacy, competition policy, regulation, and international trade. We discuss broader implications of the information revolution, examining how it changes the ways in which individuals work, play, and interact within organizations and communities. A short summary concludes the chapter.

INFORMATION TECHNOLOGY, THE DIGITAL ECONOMY, AND E-COMMERCE

Although there is no absolute agreement on what the "digital economy" is, this section provides a working definition, discusses how the term is related to e-commerce and e-business, and discusses the measurement of the digital economy, and the impacts on overall economic activity.

Defining the Digital Economy

A computer is essentially a machine for storing and processing information. Although one might count the abacus or mechanical, gear-based calculating machines as computers, the term typically refers to electronic machines that use on–off electrical signals to convey and process information. The two states, "on" and "off," based on whether an electric current is flowing or not, represent the digits 1 and 0. These are "binary digits" or "bits." Ultimately, all information that is input to a computer and is processed by it is translated into bits. *Information technology* (IT) therefore refers to anything connected to this process of storage, processing, and transmission of information converted to digital form. The use of IT for purposes related to economic transactions gives us the term *digital economy*. Here is one possible definition:

The digital economy involves conducting economic activities electronically, based on the electronic processing, storage, and communication of information, including activities that provide the enabling physical infrastructure and software.

Dramatic and rapid reductions in the costs of processing information, storing it, and sending it to others (see Table 1) have made the uses and benefits of IT potentially span the whole economy, leading to an "information revolution." On the basis of these falling costs, we have seen innovations such as personal computers, color graphics, point-and-click interfaces, and other developments that have made IT much easier to use. The Internet and the World Wide Web are the latest elements of the progress of IT over the last half-century, adding easy two-way communication of rich information (text, graphics, audio, video, etc.). The changes that the increased importance of IT brings about in people's daily lives are captured in the term "*new economy*." The term suggests that IT and the Internet shift the focus of economic activity to information, and away from traditional activities such as manufacturing. Similar terms are perhaps more descriptive: *knowledge economy*, *information economy*, and digital economy. The last of these, as noted, emphasizes the fundamental technology that drives everything: the conversion of information to digital form.

The terms "information economy" and "knowledge economy" focus on what is being digitalized. Information and knowledge are related, but distinct concepts. Information is more general and basic; it connotes anything that can be put into concrete form before digitalization. For example, a popular song is information, from an IT perspective. The sounds can be reduced to a digital form that can be stored, transmitted, and processed by various kinds of computers. If a person internalizes information about the song (its title, tune, lyrics, etc.), then that constitutes knowledge, just as the ability to write computer programs that allow users all over the world to share songs is knowledge. To push these examples further, the particular song-sharing software program is also information. In this case, knowledge helps to produce and transmit information. People can also gather information, process it in some way, and gain knowledge, as when they study how to program in a particular computer language. Some of the same distinction comes up in the differences between copyright and patent law, protecting different kinds of intellectual property rights—copyright law protects particular expressions of ideas, or information, whereas patent law protects inventions, or the ideas themselves, if they are useful knowledge. In all these cases, digitalization (through the use of IT generally, and the Internet in particular) amplifies the benefits of knowledge and makes the spread of information easier. This is one of the foundations of the digital economy.

Table 1 Falling Costs of Computing (U.S. $)

Costs of computing	1970	1999	2003
1 Mhz of processing power	7,600	0.17	0.02
1 megabyte of storage	5,260	0.17	<0.01
1 trillion bits sent	150,000	0.12	<0.01

Source: 1970 and 1999, Woodall (2000, Chart 1); 2003, author's estimates from various sources.

E-commerce

E-commerce (or *electronic commerce*) is a popular term that emphasizes the use of the Internet and associated aspects of IT for business purposes. Although businesses previously adopted IT for many internal and "back-end" activities, the Internet and World Wide Web have allowed business–consumer (*B2C*) commercial interactions to be more closely and comprehensively mediated by IT. Examples of e-commerce include buying retail items using a Web interface and paying for them by providing credit card information online; downloading media-player or other software (possibly free) over the Internet; checking the news, weather, and movie reviews on a portal, possibly "paying" for these services by giving attention to online advertisements; going to an auction Web site, and bidding on collectibles or other items; and paying a monthly subscription for Internet access, to chat online with friends or others one meets in cyberspace.

More formal definitions of e-commerce encompass all the above examples, and include commercial transactions between any kinds of organizations, not just B2C interactions. Here are two general definitions:

> Electronic commerce refers generally to all forms of transactions relating to commercial activities, including both organizations and individuals, that are based upon the processing and transmission of digitized data, including text, sound and visual images. (*Measuring Electronic Commerce*, 1997)

> In ever greater numbers, people are shopping, looking for jobs, and researching medical problems online. Businesses are moving their supply networks online, participating in and developing online marketplaces, and expanding their use of networked systems to improve a host of business processes. And new products and services are being created and integrated into the networked world. (*Digital Economy,* 2000, p. 7)

The scope of what constitutes commercial transactions is taken to be quite broad in these definitions. Information gathering or exchange that does not directly involve a direct monetary payment may still have an economic motivation. Even leisure-related activities typically require some measured economic activity. In the example of using the Internet for chatting, one pays for access to the infrastructure that enables the leisure activity. Activities that involve the government (e.g., filing one's individual tax return electronically over the Internet) are not "commercial" in the narrow sense, but are clearly related to economic activity that is measured in the national accounts statistics.

Of course, not all IT-based activities qualify as e-commerce in the sense of involving the Internet. For example, many home activities involving PCs do not involve Internet use at all: record keeping, children's homework, creating (paper) greetings cards, typing holiday newsletters, and so on. Similarly, small retail stores may have computerized inventory systems that have no links to any other computer. However, this gap between computer use and Internet use is shrinking, and for many individuals and businesses, using computers or IT automatically means using the Internet.

E-business

The encompassing definition of e-commerce presented earlier includes a broad range of online transactions and interactions that are connected to some economic motive. Therefore, this is broader than the term e-business, defined as the use of IT, including networked computing, by business firms. For example, if individuals transact directly online, so that no business firm is directly involved, then that would qualify as e-commerce, but not e-business. Similarly, we include government–individual transactions in e-commerce but not in e-business.

E-business (and therefore e-commerce) includes not just transactions across firms or individuals but also activities that take place within the boundaries of a business but do not cross them. Internal accounting, inventory control, and other forms of business record keeping and tracking have been electronically based in industrial countries for over a decade, especially in larger businesses. These purely internal records and transactions, when handled electronically, are often what e-business is taken to refer to. The use of IT provided cost advantages over traditional means (i.e., paper) in terms of storage, manipulation and retrieval of large amounts of information, provided that the scale of use was large enough to spread the substantial fixed costs initially associated with IT investments. In fact, until computers became affordable as household items as a result of falling costs and associated innovations, large organizations were the only purchasers of IT products and services.

In fact, the digital economy, in the form of business-to-business (B2B) transactions based on older electronic communication methods (electronic data interchange or EDI, using proprietary software and dedicated communication links), substantially predates the Internet and the World Wide Web. Electronic links between financial firms, and between large retailers and their suppliers, were two prominent examples of this form of e-business, or B2B e-commerce. The Internet and World Wide Web have extended the economic feasibility of such links to a much wider range of businesses, through their use of shared networks and non-proprietary communication software, and the resulting reduction in the costs of information exchange. Advances in ease of use and speed of transmission have also contributed to this trend, by further increasing accessibility and flexibility. Another potential impact of these developments, which supports the use of broader definitions of terms such as e-commerce and e-business, is the blurring of the boundaries of the firm, as information flows more freely across firms as well as within them.

SIZE, GROWTH, AND IMPACT OF THE DIGITAL ECONOMY

The size and growth of the digital economy can be gauged in several ways. Basic measures of numbers of Internet users, Web sites, and so on are popular. These do not directly measure economic activity that takes place online, though the provision of Internet access is itself an economic activity. More direct measures are figures on transactions that take place online, as well as the share of IT-related activities in the overall economy. This section examines in turn these different approaches to measuring the digital economy and its impacts.

Internet Use

Three kinds of statistics that are often used to gauge the growth of the digital economy are the number of people with Internet access, the number of unique Web pages, and the number of Web sites. The U.S. Department of Commerce, using data from the U.S. Census Bureau, reported that 143 million Americans, or 54% of the population, were using the Internet in September 2001, up from 117 million 13 months earlier (*A Nation Online,* 2002). (See http://osecnt13.osec.doc.gov/public.nsf/docs/Evans-Census-Online. Nielsen NetRatings estimated a higher number of Americans, 168 million, online in January, 2001 [NUA, 2001].) Furthermore, a broader cross-section of Americans is using the Internet, reducing fears of a "*digital divide*." In particular, the growth in Internet use during this period was fastest among Americans with household income less than $15,000 a year.

The Internet is also increasingly global, with the worldwide number of users estimated at 619 million, of whom over half used another language than English (see http://www.glreach.com/globstats/ where links to the original data sources are available). With the exception of the U.S. and a few European and Asian countries, the numbers of Internet users are still low relative to population sizes, reflecting generally lower levels of income in much of the world, but as costs continue to fall, even poor villagers in Asia, Africa, or Latin America are beginning to use the World Wide Web to get weather and crop price information or to check on village land ownership records from government Web sites. Table 2 gives a sampling of Internet use by language, from Global Reach (2003).

Table 2 Global Internet Use

Language	Internet users (millions)	Total population (millions)	Percentage of world economy
English	230.6	508	33.4
European languages (excl. English and Spanish)	176.9	868	25.0
Spanish	47.2	350	8.9
Arabic	5.5	300	1.6
Chinese	68.4	874	13.0
Japanese	61.4	125	8.0
Korean	28.3	78	2.0

The number of unique World Wide Web pages was reported to be more than one billion by January 2000, up from just 100 million in October 1997 (Inktomi, 2000, and Yahoo, 1997). The number of Web sites in May 2003 was about 40 million, up from 19,000 in August 1995 (BBC, 2002, and Netcraft, 2003). The graph of Web sites (Figure 1) is representative of the kind of growth that the Internet and World Wide Web have seen, in terms of availability and of use, with rapid early growth, slowing down dramatically, and declining slightly, in 2001–2002, although the latest figures (May 2003) indicate a resumption of growth (Netcraft, 2003).

Of course, the numbers of users or Web sites does not indicate how much time people actually spend online and how that time impacts economic activity. In particular, they do not tell us how much money people spend directly or indirectly as a result of their Internet use. For that, one would use the approach of national income and product accounting, which estimates value added in market transactions. There are problems with this way of measuring economic activity, such as the failure to account for the value of time used in nonmarket transactions, or in activities that affect market transactions, but official methods of calculating economic activity are the best we have. (For example, time spent in gathering information that affects purchase decisions is not valued in national accounts. Shifts in this activity from traditional, physical methods [such as browsing print media, telephoning, and driving around to stores] to online search will not show up in the data, except as changes in business spending [from magazine ads to Web ads], or even *reductions* in economic activity [less spending on magazines by consumers].) In particular, they are designed to capture market-based economic activity relatively well and to avoid problems such as double counting.

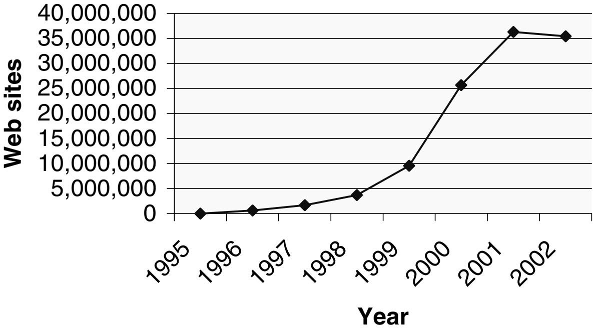


Figure 1: Growth in Web sites. Source: Constructed from data retrieved from BBC (2002) and Netcraft (2002).

Types and Measures of Online Transactions

The problem of double counting is avoided when one looks at final sales to consumers. B2C e-commerce seemed to hold out great potential in 1999 and the early part of 2000, resulting in a rather frenzied burst of entrepreneurial activity backed by eager venture capitalists. This fever has cooled, but the growth remains. The U.S. Bureau of the Census estimated electronic retail (*e-tail*) sales in 2001 to be $32.6 billion. This figure grew to about $43 billion in 2002, though even after this impressive growth, online sales remained only about 1 to 2% of all US retail sales. (These statistics are reported at http://www.census.gov/mrts/www/current.html, which also gives historical data, calculation methods and charts. Other surveys give somewhat higher figures. For example, AOL reported that its members alone spent $33 billion online in 2001 [Kane, 2002]. However, this may include travel spending, which, if excluded reduces estimates considerably. The figures reported in the text exclude travel spending, which would add another 50% to those figures if included [ComScore, 2002]. Figures for 2002 are from news reports on CNet.) In Europe, figures for online retail sales are less standardized, but one estimate for 2002 put them at about $30 billion, or comparable to the U.S. figures in magnitude. The estimate is from Forrester Research, as reported in the International Herald Tribune (Selby, 2002). Gartner, another IT research firm, estimated somewhat higher numbers (Gartner, 2002), but they seem too high compared to U.S. figures, given somewhat lower Internet penetration in Europe. Higher numbers for Europe may include travel and financial services transactions. See OECD (2002) for a discussion and detailed comparisons. Statistics for online retail sales in Asia are even less reliable, but one can estimate them to be about one fourth to one third of European figures, based on various Internet sources.

B2B transactions may involve products that are indistinguishable from consumer products (computers and office supplies, for example). The only difference is that they are sold to businesses rather than to households or consumers. However, a large segment of B2B transactions involve raw materials and intermediate products, as well as services that are specific to businesses (for example, accounting, human resource management, and, increasingly, information technology services). Although estimates of B2B e-commerce vary widely, it is widely agreed that the numbers are much more substantial than those for B2C e-commerce. The reasons have to do chiefly with the historical scale of IT, compounded by the cost-saving incentives of businesses in competitive markets. As noted in the earlier discussion of e-business, Internet-based e-commerce represents the impact of cost reductions in expanding the size of the communications network, and hence the market, from large firms to small ones, and to households.

There is a conceptual problem with most estimates of B2B e-commerce, which simply add up revenues from a variety of firms. Because B2B transactions involve intermediate products and services, aggregating revenues across businesses will involve double counting. Still, one can use the numbers, and changes in them, to get some idea of the importance of B2B e-commerce. Although the U.S. Census Bureau does not yet estimate B2B e-commerce, private forecasters do. Table 3 presents a range of estimates and forecasts, based on a combination of surveys and guesswork. (One area where differences in estimates can arise is with respect to EDI, which uses private networks and proprietary software, and has been restricted to larger firms. For example, the Boston Consulting Group estimated U.S. EDI for 1998 at $571 billion, dwarfing its estimate of Internet-based B2B e-commerce of $92 billion. At the same time, they projected EDI to grow only slowly, to $780 billion in 2003, while projecting U.S. Internet-based B2B e-commerce to be $2 trillion in 2003.) Although the numbers vary quite a bit, reflecting differences in who was surveyed, and possibly how e-commerce is defined, they are all of similar orders of magnitude, and all project substantial growth. Although actual estimates for 2002 are not available at the time of writing, they were probably substantially lower than the forecasts, given the slowdown in the global economy and in business investment (particularly IT investment). One can safely conclude that B2B e-commerce substantially exceeds B2C in size and is growing somewhat faster on average. Finally, the cost advantages of using the Internet, plus the benefits of being part of a larger network, are expected to cause a relative shift to the Internet from EDI. For example, retailer Sears Roebuck, one of the pioneers of EDI, has an EDI system that costs it about $150 per hour. Internet-based exchange with its suppliers could reduce this figure to as little as $1 an hour. (See Guy, 2000.) Small businesses, in particular, can use Internet-based e-commerce where EDI would not be economical. Furthermore, electronic marketplaces are potentially economically viable using the Internet but not with traditional EDI.

Table 3 Worldwide B2B E-commerce Estimates* ($ billion)

Source	2000	2001	2002 (forecast)	2003 (forecast)
eMarketer	278	474	823	1,409
Forrester Research	604	1,138	2,061	3,694
Gartner Group	433	919	1,929	3,632
Goldman Sachs	357	740	1,304	2,088
IDC	282	516	917	1,573

*Estimates are from surveys; source: eMarketer (2002).

One final category of e-commerce is consumer-to-consumer (*C2C*). Although firms such as eBay have entered the popular imagination through their electronic auctions for collectible or unique items, the value of C2C transactions is quite small. In fact, eBay now handles B2C and B2B transactions as well. Although its own revenues are still only in the hundreds of millions, these revenues represent only its commissions on transactions, which are therefore considerably larger in volume. Again, the total value of a transaction does not represent the economic value added by an activity. If a used item is bought and sold through a dealer, the dealer's profit is a better measure of the value created in the overall transaction, and this is what is measured in the national accounts. Thus eBay's revenues may be a good indicator of the importance of C2C e-commerce. Of course, a used item sold privately (say, in a flea market) will not show up at all in official accounting.

Information Technology and GDP

E-commerce, measured as actual transactions conducted online, represented only a small fraction of the U.S. Gross Domestic Product (GDP) of $10.2 trillion in 2001. However, overall spending on information technology (IT), a more liberal measure of the digital economy, is substantial. Although IT is more than just e-commerce, increasingly the boundaries are getting fuzzy. Networks exist within corporate walls and simultaneously are part of the larger network of the Internet. Telecommunications infrastructure that carries telephone conversations is also used for World Wide Web data. Total IT spending in the U.S. (without any double counting) now makes up about 7% of GDP, or over 700 billion dollars. See Jorgenson (2001) for detailed estimates. Figures from IT market research groups such as International Data Corporation (IDC) are a little higher. This includes hardware, software, services, and telecommunications spending. Of course, this means that there is still a substantial part of economic activity that is not directly related to information technology. However, it is a reasonable forecast that the 7% figure will increase over the next few decades. A slightly different measure, ICT value added as a percentage of business sector value added, yielded an average of close to 10% for 25 OECD countries in 2000, with the U.S. being somewhat above this average, and extremes ranging from over 15% for Ireland and Finland to about 5% for Greece and Mexico. (The 'C' refers to 'communication,' which has become increasingly digitalized, and is counted with IT in many statistical and conceptual exercises.)

One factor working against an increase in IT as a proportion of GDP is the fall in the costs of IT. The empirical regularity observed by Intel co-founder Gordon Moore, and enshrined as "Moore's Law," says that the number of transistors per microprocessor doubles every 18 months. This ability to pack more and more circuitry on tiny wafers of silicon keeps on reducing the cost of processing power. Similar factors are at work in storage and communication of information, resulting in enormous reductions in the overall cost of computing (recall Table 1). To the extent that only expenditures are measured when economic activity is calculated, some of the impact of the digital economy is being missed. For a simple example, a $2,000 home computer is many times faster than a $2,000 home computer available five years ago; it has much more storage capacity; and it can communicate much faster with other computers than was possible half a decade earlier. Even neglecting adjustments for inflation (which would mean that the $2,000 computer now is cheaper in real terms), the same amount of money spent now allows one to work more quickly and effectively, or to enjoy one's leisure more. Thus, the same spending on information technology today gives much more "bang for the buck" than five years ago.

The changes in computing go beyond having more capacity or saving time and encompass activities that were impossible in the past: online games, music listening and sharing, interactive distance learning, and so on. Again, these increased capabilities are not fully accounted for in the standard accounting of economic activity. Of course, these measurement problems have always existed. Innovation that introduces new products or improves the quality of old products has always been difficult to account for. One might argue, however, that IT has accelerated innovation and magnified the problem of underestimating the benefits of certain economic activities (Brynjolfsson & Hitt, 2000).

The problem of accounting for improvements in quality and variety goes beyond IT. If IT can be used to more effectively design new products or improve the design of existing products, then its value will be greater than is simply reflected in spending on IT itself. In other words, better, cheaper, more versatile computers make it possible to have better, cheaper, more varied cars, houses, toys and so on. This is partly what Brad DeLong (see Introduction) means when he says that IT amplifies brainpower. For example, in crash testing new cars, actually crashing a car could cost something like $60,000 each time. This is how it used to be done, with the results analyzed partly by computer. Simulating the entire crash on a computer can now instead be done for close to $100 (Woodall, 2000, p. 5).

Despite the seemingly obvious benefits of IT illustrated above, one paradox that proponents of the new economy have faced has been the lack of hard evidence for these benefits in the overall GDP data, measuring economic activity. A particular problem has been that increased investment in IT did not appear to be improving productivity in any measurable way. The conclusion of skeptics has been that much of this IT spending had no real impact. We turn to this issue next.

Information Technology, Productivity, and Growth

Much of the attention to productivity growth has been with respect to the United States, which has spent the most on IT and which had a prolonged slowdown in productivity growth in the 1970s and 1980s. Early investments in IT seemed to have no countervailing impact to reverse this slowdown. Analysis of the introduction of electric power 100 years ago (David, 2000) suggests that the benefits of innovation can take decades to appear in quantifiable form. This seems to fit with what happened in the last five years of the 20th century, when U.S. productivity growth did increase substantially, just as the penetration of PCs into homes approached 50% and as the Internet took off.

Other work on the recent U.S. experience (Gordon, 2000) suggests that the increase in productivity growth was confined to a small segment of the economy (computers and durable goods). Furthermore, the productivity boost may have been entirely the result of the prolonged economic expansion in the U.S. (productivity rises during economic booms). This skeptical view is supported by studies that find that productivity gains have been low in sectors where IT investments have been high. For example, measured productivity in banking and education actually fell from 1987 to 1997, even though these were the sectors with the highest spending on IT as a proportion of output. Possible explanations for the failure of IT investments to show up as improved productivity include the inability to account for time savings, increased outputs of public knowledge, availability of greater variety, and general improvements in the quality of products and services. Thus, some of the most important benefits of the digital economy could also be the ones that slip through the cracks in measuring economic activity.

Despite these caveats, two recent, comprehensive analyses, by Jorgenson (2001) and Stiroh (2002), suggest that IT has been an important contributor to productivity growth in the 1990s. Jorgenson directly traces this impact to the rapid fall in the prices of semiconductors and of IT products in general, especially after 1994. For 1995–1999, Jorgenson estimates that two thirds of the United States' productivity gains were the result of IT use. Stiroh goes even further, with a detailed, industry-level analysis of the U.S. He finds that the U.S. productivity revival was indeed broad-based, that much of it took place in IT-producing industries, and that industries that were IT-use-intensive also had higher productivity gains. Thus Stiroh's work appears to tilt the scales in favor of a positive assessment of the impacts of the digital economy, at least for the U.S.

Daveri (2000) also reaches positive conclusions with respect to the impact of IT on overall economic growth. His exercise includes 18 OECD countries, and his results for the U.S. are broadly similar to those of Jorgenson and Stiroh. For Canada, Japan, Australia, New Zealand, and 13 European countries, he obtains varied results, with Canada, Australia, New Zealand, the Netherlands, and the U.K. having relatively high contributions of ICTs to growth, with Italy and Spain at the other extreme. Daveri also discusses why his results are more positive than those of another cross-country study (Schreyer, 200), which uses a narrower definition of IT. Note that these studies use data for 1991–1997, and therefore miss the end of the 1990s, when significant growth in IT, as well as overall growth, took place.

IMPLICATIONS FOR MARKETS AND ORGANIZATIONS

This section discusses how the use of IT and the Internet affect the structure and outcomes of markets and the organization and strategies of firms. The consequences of the special nature of information as an economic product are also emphasized.

Information and Markets

The most basic way that the information revolution changes the economics of the marketplace is in making information about all kinds of products and services more widely available. Although basic models of the market system often take it for granted that information about products and about buyers and sellers is abundant, in practice, this is not the case. In fact, one of the virtues of the competitive market system is its ability to economize on the use of information. Textbook competitive markets cannot overcome some kinds of lack of information: for example, the quality of a product may be observable to sellers who provide it, but not to buyers. In practice, many kinds of institutions arise to overcome informational problems: brand names, warranties, consumer protection laws, and so on. Business firms and other organizations may themselves be viewed partly as a response to information problems that prevent the use of markets for all transactions (Coase, 1937). The market economy can be viewed as a scene of constantly shifting attempts to create advantages over competitors by finding opportunities for greater efficiency or satisfying wants more effectively.

In this situation, the availability of greater information about products and services may upset existing institutions, changing the relative costs and benefits of current ways of doing things. How firms organize their own internal operations and transactions can change, and how they interact with consumers can also change. New kinds of firms may arise simply to manage the new possibilities for market interaction. For example, firms may specialize in providing price or quality comparisons to consumers, in ways that were not cost-effective before. Firms may find it easier to outsource manufacturing, because they can maintain closer links with suppliers through regular information exchange. Other firms may provide combinations of services that were impossible or unlikely in the past, combining traditional media content (news, entertainment, and product information) with individual services such as auctions and communications.

From the perspective of consumers, or buyers more generally, the Internet lowers search costs by providing large amounts of product-related information "anytime, anywhere." In addition to information from sellers, buyers can also more easily access information from intermediaries that rank products or make price comparisons, or from other buyers. In consumer markets, this ability to gather information from dispersed buyers represents a major extension of "word-of-mouth" methods of sharing information.

Preliminary work on the functioning of markets online suggests that there are measurable effects of the greater availability of information. A survey by Smith, Bailey, and Brynjolfsson (2000) examines the evidence on four dimensions of price competition in B2C online markets, as compared to traditional transactions mediated through physical stores:

Price Levels: Are posted prices lower online?

Price Dispersion: Are prices of online sellers less spread out?

Price Adjustment: Do sellers adjust posted prices more finely or frequently online?

Price Sensitivity: Are buyers more responsive to price changes online?

Overall, the results of various empirical studies indicate that prices are lower online, there is less price dispersion, and prices are adjusted more often and in smaller increments. All of these conclusions are consistent with the general hypothesis that online markets are more competitive. The results on price sensitivity are mixed and may be related to factors that cannot be controlled for, such as differences in the characteristics of online consumers. The lack of evidence for perfect competition (persistence in price dispersion, for example) can be explained by the continued importance of factors such as trust and reputation and of continued switching costs associated with search that is still costly (though less so) and with investments by consumers in seller-specific information, loyalty programs, and so on.

On the other hand, the studies discussed above examine only posted price markets. Possible efficiency gains exist through the better matching of buyers and sellers in markets that otherwise relied on more costly methods of intermediation. In addition to online auctions of collectibles and other "used" items, job markets, personal-relationship matching markets, and many fragmented B2B markets appear to have taken off precisely because the Internet and Web allow buyers and sellers to match more efficiently, creating higher value matches, as well as lowering costs associated with agreeing on a price and completing a transaction.

Finally, the Internet expands markets across national boundaries and impacts international trade through its ability to increase the quantity and richness of information that is available. The kind of information that the Internet can provide does not remove the potential need for traditional face-to-face interactions, at least when a relationship is being formed. However, it reduces the cost of much of the search process, in terms of where a potential buyer might want to invest more time and money in gathering the information necessary to decide whether or not to transact. Furthermore, once a relationship is established, the ongoing costs of routine information exchange and even transactions are reduced. In this case, the products that are traded may well be traditional physical products; e-commerce does not involve transforming the products, but rather changing the nature of the search and transaction processes. Even in this case, the savings in terms of transaction costs may be substantial enough to significantly boost international trade. For small consumer items such as handicrafts, the Internet and Web provide a cost-effective way for even rural artisans to advertise globally.

Information as an Economic Product

A basic impact of the information revolution is on information itself as an economic good. The best-known example of the magnitude of this impact is what happened to the *Encyclopedia Britannica*. The *Britannica* was the premier encyclopedia, with thousands of pages of articles by renowned experts. It sold for thousands of dollars, priced to recover the cost of a well-paid direct sales force as well as the printing cost for its two dozen volumes. This business was destroyed by the ability to put a reasonably large amount of information on a single CD-ROM, sold for under a hundred dollars, or even bundled in "free" with a home computer system. CD-ROM encyclopedias were inferior in academic quality, but were "good enough" for most people, and the price was right.

The ability to store, process (including copying), and transmit large quantities of digital information at lower and lower costs is now the central characteristic of information as an economic product. A world where the marginal costs of providing information approach zero is a world where businesses that deal in information have to find new ways to provide value to consumers, ways for which they can actually charge enough to recover the costs of producing the information in the first place. In order for them to do this, information has to be bundled, it has to be personalized, and it has to be managed within a service that creates a long-term relationship with the buyer. To the extent that other products and services have also been bundled with information in the marketplace, and that this bundling changes, firms have to create new bundles of value. For example, online retailers try to provide their customers with suggestions and ideas, based on tracking the buying and browsing patterns of individual customers and those with similar interests. This is not dissimilar to the old personalized service in the local store, but can be done on a scale that was earlier impossible.

The ability to process large quantities of information in increasingly sophisticated ways is at the heart of the information revolution, extending not only to the ability to make suggestions about existing products, but also to the ability to design products in collaboration with individual customers and to do it at a large scale. *Flexible mass customization*, the ability to quickly satisfy the diverse wants of large numbers of individual consumers, is one of the possible pillars of business strategy in the digital economy.

The special nature of information (including knowledge) as an economic product also raises concerns for legal definition and enforcement of property rights, which is crucial for all market economies. In the digital economy, knowledge is increasingly important as a driver of economic activity and growth. To play this role, it must be part of the system of economic incentives. Knowledge is legally characterized as *intellectual property* (IP), but this is fundamentally different from physical property and requires its own system of legal definitions and rights. Thus, IT and the Internet amplify not only brainpower, but also the importance of the legal system that governs the economic rewards to brainpower. Furthermore, technological advances that make copying and sharing of information of all kinds incredibly quick and inexpensive are having a major impact on legal issues relating to intellectual property. We give a sample of the issues here.

Briefly, there are four areas of IP law: (1) trade secret law, which protects valuable information, not generally known, that has been kept secret by its owner; (2) trademark law, which protects words, names, and symbols used by manufacturers and businesses to identify their goods and services; (3) patent law, which protects new, useful, and "nonobvious" inventions and processes; and (4) copyright law, which protects original "works of authorship." All these concepts of intellectual property predate e-commerce by centuries, having developed along with capitalism and the industrial revolution. The information revolution provides some new challenges in this arena, especially in the area of copyright law, though patent law also has been stretched by the information revolution and the rise of e-commerce, and issues of trademarking have arisen with respect to basic online activities such as the use of hyperlinks on the World Wide Web.

Until a 1981 Supreme Court case, the United States Patent and Trademark Office was reluctant to grant patents on inventions relating to computer software. They reasoned that patents could not be granted for scientific truths or mathematical expressions of it, and viewed computer programs as mathematical algorithms, which are not patentable. In the early 1990s the courts clarified that an invention including software was patentable if the software controlled real-world processes, or numbers that represented real-world concepts. The commercial use of IT fits this description in ways that were unimaginable in 1981, and software patenting has exploded in the last few years. E-commerce software patents include broad (many argue, too broad) ideas such as "one-click buying" and "reverse auctions." In addition, the increased use of IT to govern internal business processes such as inventory control and workflow has also generated numerous software patents.

Copyright law differs from patent law, in preventing the copying of the expression of ideas, but not of ideas themselves. Therefore, copyright law does not protect against someone stealing an invention or someone else independently creating a similar expression. However, copyright does provide some protection against "nonliteral infringement," such as the near duplication of screen displays. The primacy of information products, or "content," in e-commerce, and the ease with which digital information can be copied and distributed have made copyright law for the Internet a major area of concern.

Market Structure and Strategy

The nature of information, where marginal costs of delivery are small and fixed costs of production may still be large, is often alleged to favor *winner-take-all* outcomes. This is reinforced by the benefits of creating and controlling large networks: consumers will presumably join the network that is already large, to get the highest benefits of being able to select a transaction partner from a bigger pool. Thus buyers will look on eBay, because it is the largest online auction site, whereas sellers will list there for the same reason. The size advantage associated with these *network externalities* keeps reinforcing itself. On the other hand, if individuals can simultaneously participate in more than one network, and if a smaller, competing network can offer a price break to attract them, then the advantage of the winner will be limited, and taking all may not be feasible.

Focusing on information leads to a different emphasis in terms of the economics of business strategy. Pure price competition has to be less important than competition along an array of different dimensions, because price competition in a world of high fixed costs and low marginal costs will lead to firms losing money. For example, the Bertrand model of price competition suggests that firms with homogeneous products will be pushed toward pricing at marginal cost, making it impossible to recover their sunk fixed costs. Pricing itself becomes more complex, with greater possibilities for differentiating across consumers and over time, and is joined by marketing, advertising, product differentiation, and ways of raising the costs for customers to switch to competitors. All these dimensions of strategy exist independent of the Internet and e-commerce, but they become more salient in a world where information technology operates throughout the stages of production (what business strategists call the "value chain"), as well as in the interaction of buyers and sellers in the marketplace. The ability to gather information about buyers and sellers, to organize this information, and to analyze it creates the potential to integrate the different dimensions of strategy in ways that were not possible earlier.

One can distinguish between two separate aspects of firms' strategies, and how they can change as a result of information technology and the Internet. First, firms can create more value, by meeting consumer wants more effectively. Thus, being able to elicit consumer preferences more directly and at lower cost potentially allows firms to design products and services that create greater consumer surplus. The design, manufacture, and delivery (in the case of digital products such as content or software) of these products can also be more efficient, through the use of IT within the firm and in its communications with its suppliers (its "supply chain"). Furthermore, consumer tastes can not only be better matched on average, but also through greater customization, where products and services can be designed and manufactured to precisely meet individual tastes, as specified by the customer online. This customization is particularly valuable for information products, such as content, and also for very personal items such as cosmetics and beauty products. Again, the use of IT internally and through the supply chain allows the more flexible approach to production required by mass customization. Where the products and services are digital, product differentiation can also include bundling in new ways. For example, online portals provide bundles of content and services that are not matched by traditional media companies or content aggregators such as newspapers and magazines.

The second aspect of business strategy is value capture. In this case, the greater ability to gather information about buyer preferences through tracking online behavior—including observational and information-gathering behavior as well as buying patterns—allows pricing to be more finely tuned, so that prices are closer to buyers' maximum willingness to pay. Thus, the ability to potentially use online interactions to gather and analyze buyer information allows greater value capture through various kinds of price discrimination. In such cases, product differentiation may be an important supporting strategy for value capture, in addition to its role in value creation. Product differentiation can reduce the ability of competitors to undercut a price discrimination strategy. The ability to vary prices more easily and often in online markets may also be used as part of a price discrimination strategy, just as seasonal sales are used in traditional retailing. Bundling of products is another example of how firms may tailor products to capture greater value from buyers, though it is not always used for this purpose.

A slightly different way of analyzing business strategies in e-commerce is to think of online shopping as offering a bundle of three different categories of goods and services: the products themselves, the service of time in physically assembling the order and delivering it, and an information service that is made possible by the infrastructure of the Internet and the World Wide Web. The digital information processing and communication capability of this infrastructure is what makes the bundling of the other two services economical. In the case of many physical products, such as groceries, the service of assembling the order and delivering it would not be cost-effective without it. The nature of the information service is what clearly distinguishes catalog shopping for physical products from online shopping. In the case of shopping in a store for physical products, the bundle offered to the buyer is more different, because in store shopping the buyer bears the costs of "last-mile" fulfillment, i.e., bringing the product home. In the case of digital products and services, the differences are greatest as compared to traditional methods, because the product itself can be delivered over the same infrastructure that is used for ordering. As one might expect, online transactions for digital products and services provide greater potential for changing ways of doing business. In the case of physical products, online transactions must still rely on physical delivery using conventional transportation methods.

Changes in Firm Organization

The use of digital technologies has several different possible implications for the organization of firms. A simple hypothesis is based on the transaction cost theory of the firm (Coase, 1937), which argues that firms exist to overcome transaction costs associated with market exchange. If using IT and the Internet reduces market transaction costs, by increasing the speed and richness of information flows, then firms will be replaced by markets. Although this argument has merit, the use of IT within firms also improves, so the opposite case can be made, that firms are able to become larger as a result of information technology. Certainly, the globalization of firms in areas such as sourcing inputs, or serving different geographic markets, appears to be aided by the use of IT, and the Internet in particular.

One reason that firms will not disappear is that factors such as the need to control complementary assets, especially different kinds of skilled labor or human capital, are a reason for the existence of firms that is not removed by improved information flows. In general, one can argue that efficient incentive provision—for current effort as well as investment-related activities—often requires the use of hierarchical organizational forms, rather than pure reliance on market transactions.

Nevertheless, two trends are evident, and both appear to be accelerated by the developments in IT and communications. First, the growth of markets, which has been driven by lower transportation costs and trade barriers and by higher incomes, as well as dramatically improved communications, permits greater outsourcing, and therefore some decrease in vertical integration of firms (Brynjolfsson & Hitt, 2000, p. 36). This is an example of Adam Smith's classic maxim that the division of labor is limited by the extent of the market. Contract manufacturing is an important example of this phenomenon, where the supply chain becomes more geographically dispersed as well as being more divided among different firms.

Second, even when they do not outsource, firms themselves are becoming more decentralized, as they incorporate IT into their internal business processes. For example, Brynjolfsson and Hitt (2000) summarize a study of large firms that found that greater internal levels of IT use were associated with "increased delegation of authority to individuals and teams, greater levels of skill and education in the workforce, and greater emphasis on pre-employment screening for education and training" (p. 35). Other studies indicate that the IT investments of decentralized firms are more productive and have such firms have higher market values, suggesting that market pressures will favor greater decentralization in the long run.

From the Chandlerian perspective of scale and scope (Chandler, 1990), the above factors pull in different directions. Firms can become more specialized, because their ability to serve geographically dispersed markets is enhanced by the Internet. For example, niche retailers can potentially sell globally, as their marginal costs of reaching new customers are substantially reduced. Trading off scope for scale means that the overall impact on firm size is indeterminate. In general, however, the lower cost of entry into online business—a Web site is cheaper to set up and operate than a physical store—suggests that smaller firms will thrive. At the other extreme, large firms can expand their scale and scope more easily using online presence and interactions, and especially in the case of digital products and services, offer very large ranges of offerings. Therefore one plausible prediction is that the size distribution of firms will become more spread out, with firms in the middle losing out to those at either end (e.g., *The Economist,* 2002).

Returning to globalization, an important consequence of digitalization is the ability to deliver digital products and services across national boundaries, using the Internet. A related issue is the impact of greater information availability, which expands market reach. In both cases, the ability to complete transactions online is an additional feature, though not an essential one. Skills provision and financial services provision are examples of electronic delivery across national boundaries. Small IT projects may be handled entirely online. In other cases, there is typically some face-to-face meeting (high-bandwidth information exchange) and agreement, followed by more routine exchanges or deliveries of services that take place online. The outsourcing of software development from the developed world to countries such as India and Israel represents an important example of such activities. In another example, retail financial services may be conducted entirely online across national boundaries. An investor in Europe trading in U.S. stocks can fill out the application forms, transfer money into a U.S. account, and trade without using paper or moving from his or her screen. Commissions earned by the online brokerage then represent a payment for services that are international.

A final aspect of changes in firm organization in the digital economy is related to the new kinds of products, services, and delivery methods that are possible. In many cases, new types of digital economy firms are essentially new intermediaries, providing expertise or reputation, or economizing on transaction costs. The following classification of firms is in terms of how they combine information, time services, and physical goods and services in new ways in their offerings to customers.

Information request services provide general information on demand through search engine technology, a very basic aspect of the World Wide Web.

Content providers package particular sets of content, rather than enabling general searches for any kind of information, and represent a new kind of media firm.

E-tailers are a carryover from the world of physical retailing, including catalog sales, offering a different bundle of physical products, time, and information services than traditional retailers.

Exchanges and brokers operate electronically without physically bringing together buyers and sellers or the objects being sold and act as "market-makers" or "market-expanders."

Community creators provide online mechanisms for communication and for collaboration.

Infomediaries focus only on providing information pertaining to potential market opportunities and are potentially neutral with respect to buyers and sellers.

Portals are aggregators, or diversified firms, combining the six types of firms above.

Infrastructure providers make and sell the hardware, software, and services that allows other firms to process, store, and send the information that makes their businesses possible. These include communications equipment, Web hosting equipment, connectivity services, hosting services, and many kinds of software developers.

As the seventh category itself suggests, there are many overlaps and combination possible in these functions within different digital economy firms.

GOVERNMENT POLICIES

Any market economy relies on government regulation to maintain a framework of laws and property rights that allow production and exchange activities to occur in a stable environment. The rise of the digital economy has several impacts on government management of the economy and the legal environment within which it functions. We briefly examine several of these impacts in this section.

Contracts

Traditional commercial transactions are governed by well-defined laws and legal precedents. In particular, there is a clear concept of what constitutes a legally binding contract. Paper documents with signatures are the norm for contracts. Sometimes, notarization to authenticate the signatures is required. There are also disclosure requirements and escape clauses, particularly for consumers transacting with businesses. The legal issues in e-commerce contracting revolve around how identities can be verified, signatures can be authenticated, and content can be protected when information is stored, processed, and transmitted electronically.

Security of content is provided by encrypting (encoding or scrambling) that content. Encryption is an old idea, but information technology permits the use of more powerful mathematical algorithms and therefore more secure encryption. Security is different from integrity of content. Authentication of content integrity and of the sender's identity use mathematical ideas and technology similar to those for encryption. In physical markets, checking IDs or signatures is a well-established procedure. Even over the telephone, ID can be checked by providing certain information that authenticates identity (the last four digits of a social security number, for example). Digital signatures achieve similar goals for electronic communications: they can identify the sender and also authenticate content.

The growth of e-commerce itself will depend on the ability of two parties to complete a contract, sign it in a legally binding manner, and transmit it, all purely electronically. The technology is not the stumbling block to this goal. The issue is one of clear, generally agreed-on legal standards. For example, in the U.S., in June 2000, the President signed (electronically as well as with the traditional pen) a bill that sets these standards and will make it possible for businesses to close deals with electronic contracts and digital signatures. Similar legislation has been passed in other industrial countries.

Electronic contracts are especially attractive for B2B transactions. However, the possibility of electronic contracting will probably require some updating of rules that protect consumers. Because a large percentage of households are still not online, presumably consumers should still have the right to have all contract details and subsequent pertinent notices on paper, without financial penalty. This makes the cost saving that electronic dealings offer to businesses harder to achieve, but presumably these will come with time, as electronic communication become cheaper and more ubiquitous. The technology of digital certification will also have to become more widespread and widely understood for it to serve the everyday needs of B2C transactions.

Intellectual Property

Phenomena such as the widespread copying of digital music, using file-sharing software available from many different commercial and noncommercial providers, have heightened concerns about enforcement of copyrights on the Internet. Some have called for more stringent copyright laws, and, in the U.S., the Digital Millennium Copyright Act (DMCA) of 1998 did introduce some additional protections in the guise of updating previous law to cover new technologies. However, it may be that existing laws are quite sufficient, as court rulings against Napster and others suggest, as long as the interpretation of the laws is strong enough.

The DMCA included penalties for cracking protections designed to protect unauthorized copying. However, the legal application of the DMCA has had an effect on free speech, with examples such as a professor of computer science forgoing presentation of a research paper outlining methods of overcoming copy protection, in the face of a threatened lawsuit. Lobbying by industry groups that have ownership interests in copyright has also motivated potential U.S. legislation that requires copy protection to be hardwired into consumer electronics items. In such cases, the doctrine of "fair use" in copyright law also appears to come under attack. In the U.S., the DMCA was complemented by another law, extending copyright protections by 20 years. Although European copyright laws have not tilted so much against users, the U.S. is a global leader in the production of copyrighted material such as music and films.

In the arena of patents, it has been court interpretations rather than new legislation that have increased the scope of patent law, with hundreds of relatively broad software patents being granted in recent years. It has been suggested that a new category of patents, with a shorter lifetime, be granted for software, but this is likely to create further problems for assessing patent applications. It also misses the real problem, which is that of inadequate resources in the U.S. PTO. Again, the U.S. is the largest market for intellectual property, and U.S. patent rules are disproportionately significant in the global context. In general, the apparent broadening of both copyright and patent protection can be seen as a response to a situation where intellectual property is increasingly important for creating economic value, but also easier to copy or imitate. Thus, it is a symptom of the digital, information, or knowledge economy.

Privacy

The digital economy, with its greater flows and tracking of information, raises serious concerns about privacy. Information about consumers allows firms to increase profits through various kinds of price discrimination. At the same time, some consumer information can help firms to tailor their products more effectively to consumer preferences. Privacy concerns often center on how the information is collected—do consumers realize that their behavior online is being tracked, or that cookie files are being deposited on their computers?

Further issues arise with respect to who else may properly see the information collected. A customer may not mind a seller tracking his or her buying habits in order to serve him or her better, but may not want the firm to sell that information to other firms. A related issue is the use of such information for mass marketing e-mail, commonly known as "spam." Employees, too, may find that their electronic communication and Internet browsing from work can be monitored by employers with great precision and intrusiveness. Finally, there are all kinds of information that various public agencies collect. Often such agencies are required to make that information available in response to requests from members of the public. However, making that information available online makes access much easier and broader than other forms of availability, with possible negative consequences.

The U.S. Congress was actively considering Internet privacy legislation after a report from the Federal Trade Commission in mid-2000 indicated that self-regulation was not working uniformly, with some Web sites proving resistant to privacy concerns. One stumbling block for legislative agreement was the simple issue of whether businesses should be required to explicitly get consumers to "opt in" to allow their personal data to be used beyond the specific transaction or relationship, or whether the burden should be on consumers to explicitly "opt out." Businesses naturally favored the latter approach, which gave them much more leeway. Business-supported groups have tried to argue that privacy legislation would be inordinately costly, and also that consumers do not care enough for it to be worthwhile. On the other side are groups such as the American Civil Liberties Union and Consumers Union that want stronger safeguards against data-collection practices that do not involve explicit consent.

The aftermath of September 11 has tilted the scales against strengthening of privacy, because security has become a much greater concern. Therefore, it seems unlikely that any meaningful Internet privacy legislation will be passed any time soon in the U.S., leaving its online privacy laws some way behind those of the European Union, which protect consumer privacy more stringently.

Antitrust and Regulation

Antitrust laws are designed to prevent monopolization of industries, as well as anticompetitive practices such as price fixing. Does the digital economy require modifications in the government's enforcement of antitrust policy, or even a change in the antitrust laws themselves? There are three key areas in which the proponents of a modified approach to antitrust make their points. First, there is the argument that antitrust enforcement must account for the impacts on future innovation. The second argument is that network externalities and the economies of scale associated with information goods make monopolies more likely or more natural ("winner takes all"), and hence they must be tolerated—otherwise there will be no market or unnecessarily high costs. The third argument is that complementarities in information goods require firms to cooperate in ways that might seem collusive by more traditional measures.

Considering the first of these, the increased importance of technological innovation, and of patenting, certainly makes these variables more important in a firm's business strategy, but it does not, by itself, imply that antitrust law has to change. Firms can profitably innovate, using patent law, without having to run afoul of antitrust law. Second, network effects are demand-side economies of scale, which can interact with the usual cost-side economies of scale to promote market dominance. If information goods are subject to both kinds of economies of scale, one might have to be resigned to more case of "natural" monopoly, driven purely by the structural characteristics of the market, rather than any illegal behavior. However, the importance of such natural monopolies is probably overstated, and their persistence is unlikely if the protection for patents is not too stringent. (Arguments relating to network externalities as well as to innovation have been made in the Microsoft antitrust investigations, both of which concluded with out-of-court settlements. One can argue that any monopoly that Microsoft might have is related more to traditional anti-competitive practices such as the nature of its contracts with distributors, rather than any special features of the digital economy. In the European Union, there is an ongoing antitrust investigation of Microsoft.) Finally, technology goods have to work in systems and are characterized by strong complementarities. This often requires firms to collaborate in research and development, as well as in production and installation. However, as long as single firms, or firms acting together, do not engage in behavior that reduces competition or harms competitors, there is no violation of antitrust laws. What is needed is not a reform of the laws, but simply enforcement by government officials who understand technology well enough to sort out different kinds of cooperative behavior among technology-oriented firms.

Laws to manage privacy issues and antitrust laws can both be considered as major examples of regulation of private economic activity by the government. They are not the only ones. There are specialized regulations for different sectors of the economy, such as financial services and telecommunications. There are also regulations meant to protect certain groups, or to control certain types of activity. For example, pornography and hate materials are controlled, and gambling is a heavily regulated activity. All these forms of regulation are affected by the Internet. Much of the problem is simply in the freedom with which information can be disseminated and shared on the Internet. The location of activities is also a problem: for example, online gambling can escape controls that are designed to operate within geographic jurisdictions. Controls on forms of payment, using the major credit card companies, for instance, can be a way to solve the jurisdictional problem. The credit card companies cannot afford to be partners in crime. Here the law needs to change to deal with ability of digital activities to escape the requirement of meeting in a particular place.

In the case of financial services, the issues have to do with the quantity and the veracity of information that is made available. The Internet makes scams easier to implement in some ways, but the basic laws do not need to adjust. In the case of telecoms, regulatory issues are centered on the technological changes that digitalization has introduced, making more effective competition possible. The U.S. Telecom Act of 1996 began the process of moving regulation into the modern era. Some regulation is still needed because parts of the network are still potential monopolies. Local telephone companies—the so-called Baby Bells—in particular, have maintained their strongholds. Regulations to allow interconnection to parts of the Baby Bell networks by competing carriers have not really enabled the latter to gain significant market shares. While protecting their traditional markets in voice communications, the Baby Bells have been lobbying for the ability to compete more freely in markets for data communications, that is, the underpinnings of the Internet.

International Trade

Countries often have customs duties or tariffs on imports, and these clearly affect international trade. They may also use quantitative restrictions on the entry of certain goods and services. An extreme case of this would be a total ban. Various reasons for restricting international trade do exist: government revenue raising, control of undesirable materials, protection of some domestic groups, and so on. Individual country choices made without coordination may lead to outcomes that are worse for all countries. Therefore, in order to try to achieve some measure of cooperation that can improve outcomes, trading nations use the World Trade Organization (WTO) to frame and enforce rules for international trade. Having such an organization does not remove conflicts, but it provides a mechanism for more orderly handling of disputes, as well as a clear set of "rules of the game."

The current provisional WTO agreement is that trade restrictions should not apply to electronic transmissions over the Internet. In the example of a European purchasing U.S. stocks or a U.S. hotel room while on vacation, this leads to a symmetric treatment of online and offline transactions. In other cases, however, there is a difference. Thus, purchasing a large number of music CDs from another country might be subject to a customs duty (small purchases from abroad are exempt in the U.S., though not in many countries), but obtaining the same quantity of music as electronic files would escape the import tariff. This is superficially similar to the issue of sales taxation within the U.S., but here we are looking across national borders, and in the U.S. sales tax case, the tax must be paid if the transaction is in-state rather than across states. Hence the two cases are somewhat different, though broadly related in spirit.

A further issue with respect to information products in particular is the blurred line between goods and services. For example, software development is a service that is now offered across national boundaries. Also, software is typically licensed rather than sold, and leasing software is common. Because traditional services, such as those in the financial sector, have been treated under a separate set of rules (the General Agreement on Trade in Services, or GATS), which is newer and more restrictive than the rules for goods or products, there is disagreement as to which set of rules should govern software. Countries such as the United States, which are producers and net exporters of information and related services, argue for the application of the rules for conventional trade, rather than the GATS.

WORK, PLAY, AND COMMUNITIES

Work is a large fraction of our lives. It is useful to recognize how drastically work was altered by the industrial revolution, the introduction of factories, and the rise of large corporations. Cottage or home production became relatively insignificant, as mechanization and economies of scale caused work to be concentrated in factories and offices. Now IT has loosened the bonds of location, making work once again more flexible for many.

Location of Work

Several trends have driven the changes in work. First, the increase in the importance of services relative to manufacturing, and of the information economy in general, reduced the proportion of factory jobs. Next, the falling cost of computing power allowed many tasks to potentially be performed at home, rather than in the office. Most importantly, the Internet has removed the isolation of the home worker. Communication and collaboration can take place among workers in different locations. Physical proximity for many jobs becomes only a part-time requirement.

Some of the change is just in freedom of location. However, as we discussed in a previous section, some of the change is in the nature of the firm itself, sometimes reducing the bonds that define a firm. Employees in such cases can become independent contractors, with their own capital (human and physical), almost harking back to the preindustrial era of home production.

Another change that comes from the falling cost and increasing versatility of communication over the Internet is in the global distribution of work. The customer in the U.S. may have a telephone query answered by someone in Ireland, India, or the Philippines. Computer programming or program testing assignments may be sent over the Internet wherever people with those skills are available, to be completed and sent back the same way. The supply of some kinds of skills becomes global rather than local or national.

There is also a time dimension to this geographical dispersion of work. Time differences across the globe allow 24-hour customer service to be more cost-effective. In areas such as software development, they also allow global shift work. For example, two project teams in the U.S. and India can collaborate to achieve an almost continuous workflow, utilizing the night-and-day time difference between the two countries. However, the ubiquity of digital communication devices means that the notion of times and places where work does not intrude is severely eroded. A knowledge worker may find she is expected to access her e-mail via a wireless handheld computer at home, on vacation, and in general outside the normal place and time of work.

Leisure Activities

Leisure activities in the industrial age have been shaped by scale and specialization, just as happened in the case of work. Sports and the performing arts have become large-scale spectator or audience events. Radio and television introduced broadcasting, creating mass markets for entertainment while removing locational barriers. Recording technologies expanded the scope for listening to music or watching movies, while introducing greater choice into consumer decisions. All these developments in people's leisure activities are enhanced and broadened by the Internet. Inexpensive digital recording and transmission of music and video provide a range of options unimaginable in the past.

Perhaps the greatest impact of the substitution of bandwidth for being in the same place has been in game playing. One can play traditional games, such as bridge and chess, over the Internet, with opponents and partners who may be anywhere in the world. More widespread is the enormous expansion of online game playing. Computer games become virtual worlds where individuals act out their fantasies and try out strategies. Game characters take on lives of their own, becoming valuable commodities themselves for game players who want to win any way they can. At one level, the interaction is no different from that of board games that have been played for hundreds or thousands of years, that of stylized competition. However, the complexity of such games has increased exponentially, and the Internet has demolished distance in creating communities of game players.

Just as IT has allowed work to intrude into leisure spaces, it has also allowed the reverse phenomenon to take place. Workers who sit at a computer may play solitary games. If they have Internet access, they may engage in all kinds of leisure activities, including browsing news and entertainment content, shopping, and chatting as well as game playing. Hence, employers may respond with new kinds of monitoring and restrictions, as was discussed earlier in the context of privacy.

Online Communities

From online game players to members of a project team designing a software program, people at work and play form communities based on shared goals or interests. Information technology allows these communities to be freed from the need to share a physical space. Interactions take place on computer screens instead of face-to-face, but the interactions that are possible in cyberspace are getting richer and richer, allowing more and more communities to form. In particular, work collaborations of increasing complexity are becoming possible on the Internet, with simultaneous or asynchronous participation in activities involving product research, design, and development.

Work and play are not the only glue that binds communities. Any kind of shared interest can provide the impetus. Sufferers from a disease, fans of a rock star, or collectors of sports memorabilia can join together to exchange information, ideas, and experiences over the Internet. These communities may provide commercial opportunities, because they provide access to that ever-scarcer commodity, "attention," but they may also lead to more profound social changes. Political organization, in particular, takes on a new dimension, perhaps expanding the scope of democracy, while definitely changing its nature. Possibilities exist for Internet-based comment, feedback, and even voting by citizens in communication with their governments.

Perhaps the most remarkable change of all is how, in just a few short years, the majority of people in the industrialized world have come to take for granted so many possibilities that alter their lives and may reshape the social fabric. The Internet, at its core, is a very human-centered development. This may seem somewhat paradoxical. The underlying information technology is complex and abstract. But the Internet and its associated technologies allow people to be creative, to express their individuality, and to communicate and connect with other individuals in new ways and with new freedom. This extension of basic human capabilities, amplifying humanity and not just brainpower, is why the Internet excites so many and inspires sometimes-overstated rhetoric.

CONCLUSION

The digital economy is much more than simply online shopping. It involves a fundamental transition that has been taking place for over two decades and that is based on the rapidly falling costs of processing, storing, and transmitting information in digital electronic form. Some of this transition was obscured by the dot-com mania, which often focused on using the Internet as a marketing and retailing channel. In fact, the digital economy includes this as just a small part. The internal organization of firms is changing, their nature is changing, the kinds of interactions that are possible between different economic agents are changing, and the locations of different activities are changing. Measures of the digital economy can understate or overstate the current impacts, but they do seem to emerge in recent academic work, and the impacts appear to be increasing.

GLOSSARY

Bandwidth Most often informally used to refer to the speed with which digital data can be transferred over a specific connection (telephone wire, cable, optical fiber, or wireless). It also has a more precise technical meaning, with similar implications, e.g., 10 Mbps, or 10 megabits per second.

Cyberspace All electronic interactions and data, especially those that are mediated by the Internet. The term was coined by William Gibson in his science fiction novel *Neuromancer*.

Digital divide A situation where particular socioeconomic groups have access to the Internet and information technology at levels that are substantially higher than other groups.

Digital economy A term that emphasizes the importance for the overall economy of information that is stored, processed, and exchanged in digital electronic or optical formats.

E-business A subset of e-commerce, including all electronically aided transactions and activities of businesses, among which are internal accounting, inventory control, and communications.

E-commerce Short form of electronic commerce; refers to doing business electronically, based on the electronic processing, storage, and communication of information, and includes activities that provide the enabling physical infrastructure and software.

E-commerce interaction types: B2B, B2C, C2G, C2C, B2E Acronyms for different interactions, implicitly, but not necessarily, electronic, between businesses, consumers, governments, and employees.

EDI Electronic data interchange—the use of proprietary software and leased telecommunications lines for communication between firms, typically at different points of the value chain.

E-tailing Electronic retailing.

Flexible mass customization The ability to quickly satisfy the diverse wants of large numbers of individual consumers.

Information economy A term that emphasizes the importance for the overall economy of all kinds of information, including entertainment, news, market and business information, research, and personal communication.

Information revolution A term that emphasizes the dramatic effects of the steep fall in the costs of processing, storing, and communicating information as a result of advances in information technology.

Information technology Any aspect of technology, including hardware, software and services, that involves data in digital electronic or optical formats, including technologies for processing, storing, and transmitting such data.

Intellectual property Useful inventions, original expressions of ideas, and names or symbols used in business, the ownership of which is protected by various categories of law (trade secret, patent, copyright, and trademark).

Knowledge economy A term that emphasizes the importance for the overall economy of all kinds of knowledge, including various types of expertise, skills, and understanding of particular markets, with an implicit emphasis on mathematics, science, and technology.

Moore's Law An empirical regularity, described by Intel co-founder Gordon Moore, that the processing power of microprocessors doubles every 18 months; therefore an indicator of the rapid pace of innovation in the digital economy.

Network externalities A situation where the value of being part of a network depends on the number of other members of a network—typically this value is positive. For example, if the value of a network with n members to an individual depends on the number of possible connections in the network, it is proportional to $n(n-1)/2$ ("Metcalf's Law"), and the marginal value of an additional member is proportional to n, the network size.

New economy A term that encompasses the ideas behind the terms "digital," "information," and "knowledge" economy, but also sometimes connotes that the working of the economy is changed, either because information is a good with high fixed and low marginal costs, so competition is less stable, or because faster information flows reduce adjustment times and hence swings in the economy.

Online Being actively connected to, or being a user of the Internet (and possibly other electronic networks).

Supply chain The portion of a firm's value chain that involves its suppliers.

Value chain A schematic representation of a firm's stages of production, possibly including activities that take place upstream or downstream of the firm's own activities. Examples of value chain stages include inbound logistics, production operations, outbound logistics, marketing and sales, and after-sales support.

Winner-take-all A market situation where the leader dominates because high fixed costs and low marginal costs of producing information favor one or a few large firms, or because users of a network get much higher benefits when the network is larger.

CROSS REFERENCES

See *Click-and-Brick Electronic Commerce; Digital Divide; Electronic Commerce and Electronic Business; Global Issues; Legal, Social and Ethical Issues.*

REFERENCES

A nation online: How Americans are expanding their use of the Internet (2002, February). Washington, DC: U.S. Department of Commerce, Economics, and Statistics Administration and National Telecommunications and Information Administration.

BBC (2002, January 2). Internet starts to shrink. Retrieved May, 28, 2003, from http://news.bbc.co.uk/1/hi/sci/tech/1738496.stm

Brynjolfsson, E., & Hitt, L. M. (2000, Fall). Beyond computation: Information technology, organizational transformation and business performance. *Journal of Economic Perspectives, 14*(4), 23–48.

Chandler, A. (1990). *Scale and scope: The dynamics of industrial capitalism*. Cambridge: Belknap Press.

Coase, R. (1937). The nature of the firm. *Economica, 4,* 386–392.

ComScore (2002, January 16). U.S. online consumer sales surge to $53 billion in 2001, ComScore reports. Press release. Retrieved May 28, 2003, from http://www.comscore.com/press/release.asp?id=61

Daveri, F. (2000, September). *Is growth an information technology story in Europe Too?* University di Parma and IGIIER working paper.

David, P. A. (2000). Understanding digital technology's evolution and the path of measured productivity growth: Present and future in the mirror of the past. In Erik Brynjolfsson & Brian Kahin (Eds.), *Understanding the digital economy* (pp. 49–98). Cambridge, MA: MIT Press.

Digital Economy 2000 (2000). U.S. Department of Commerce. Retrieved April 30, 2003, from http://www.stat-usa.gov/pub.nsf/vwNoteIDLookup/NT00002282/$File/digital2000.pdf

eMarketer (2002). E-commerce trade and B2B exchanges. Retrieved May 28, 2003, http://emarketer.com/products/report.php?2000091

Gartner (2002). GartnerG2 says European online shopping market will reach 97.8bn in 2002. Retrieved May 1, 2003, from http://www.gartnerg2.com/press/pr2002-03-19b.asp

Global Reach (2003). Global Internet statistics (by language). Retrieved May 27, 2003, from http://www.glreach.com/globstats/

Gordon, R. J. (2000, Fall). Does the "new economy" measure up to the great inventions of the past? *Journal of Economic Perspectives, 14*(4), 49–74.

Guy, S. (2000, February 29). Sears, French giant in online venture. *Chicago Sun-Times*.

Inktomi (2000, January 18). Inktomi WebMap: Press release. Retrieved May 28, 2003, from http://www.onlinemag.net/OL 2000/engine5.html

Jorgenson, D. W. (2001). Information technology and the U.S. economy. *American Economic Review, 91*(1), 1–32.

Kane, M. (2002, January 2). AOL members set shopping record. Retrieved May 1, 2003, from http://news.com.com/2100-1017-800049.html

Measuring electronic commerce (1997). Paris, France: Committee for Information, Computer and Communications Policy, Organisation for Economic Co-operation and Development.

Measuring the information economy (2002). Paris, France: Organisation for Economic Co-operation and Development. Retrieved May 28, 2003, from http://www.oecd.org/pdf/M00036000/M00036089.pdf

Netcraft (2002). Netcraft Web server survey archive. Retrieved May 28, 2003, from http://www.netcraft.com/survey/archive.html

Netcraft (2003). May 2003 Web server survey. Retrieved May 28, 2003, from http://news.netcraft.com/archives/web_server_survey.html

NUA (2001). Nielsen NetRatings: Sixty percent of Americans are online. Retrieved May 1, 2003, from http://www.nua.ie/surveys/index.cgi?f=VS&art_id=905356461&rel=true

Schreyer, P. (2000, March 23). *The contribution of information and communication technology to output growth: A study of the G7 countries*. Paris, France: OECD, Directorate for Science, Technology and Industry, DSTI/DOC. STI Working Paper 2000/2. Retrieved May 28, 2003, from http://www.oecd.org/pdf/M00000000/M00000383.pdf

Selby, N. (2002, November 25). E-tailers are primed for happier holiday. Retrieved May 1, 2003, from http://www.iht.com/articles/78016.html

Smith, M. D., Bailey, J. P., & Brynjolfsson, E. (2000). Understanding digital markets: Review and assessment. In E. Brynjolfsson & B. Kahin (Eds.), *Understanding the digital economy: Data, tools, and research* (pp. 99–136). Cambridge, MA: MIT Press.

Stiroh, K. J. (2002). Information technology and the U.S. productivity revival: What do the industry data say? *American Economic Review, 92*(5), 1559–1576.

The Economist (2002, February 2–8). The real-time economy: Re-engineering in real time.

Woodall, P. (2000, September 23). The new economy. *The Economist*.

Yahoo (1997, October 3). Internet volume doubling every 90 days. Retrieved May 28, 2003, from http://www.nua.com/surveys/index.cgi?f=VS&art_id=875893545&rel=true

FURTHER READING

ComScore: An Internet marketing and audience measurement company. Retrieved April 30, 2003, from http://www.comscore.com/

Emarketer.com: E-business research source. Retrieved April 30, 2003, from http://www.emarketer.com/

Jupiter Research: An Internet marketing and audience measurement company. Retrieved April 30, 2003, from http://www.jupiterresearch.com/

Netcraft: An Internet services company that provides data on the number of Internet servers. Retrieved April 30, 2003, from http://www.netcraft.com/

Nielsen//NetRatings: An Internet audience measurement company. Retrieved April 30, 2003, from http://www.nielsen-netratings.com/

NUA Surveys: Resource for Internet trends and statistics. Retrieved April 30, 2003, from http://www.nua.com/surveys/

Organisation for Economic Co-operation and Development: E-commerce and information technology statistics and reports. Retrieved April 30, 2003, from http://www.oecd.org/EN/home/0,,EN-home-29-nodirectorate-no-no–29,00.html

STAT-USA/Internet: A service of the U.S. Department of Commerce, is a site for the U.S. business, economic and trade community, providing authoritative information from the Federal government. Retrieved April 30, 2003, from http://www.stat-usa.gov/

Wyckoff, A. (1999). *The economic and social impact of electronic commerce: Preliminary findings and research agenda*. Paris, France: Organisation for Economic Co-operation and Development.

Digital Identity

Drummond Reed, *OneName Corporation*
Jerry Kindall, *Epok Inc.*

WHAT IS DIGITAL IDENTITY?

Digital identity means many things to many people. In general, the term refers to the class of technologies, standards, services, and applications that enable a real-world identity to be represented digitally on a network and that manage the data related to the identity (its *attributes*) and control access to various network resources. Technically, any addressable resource on a network has identity; to many people, though, the term is most meaningful when describing how an individual user's personal identity and personally identifiable data are modeled, represented, controlled, and shared on the network.

Digital identity touches on issues of trust, privacy, security, and interoperability that are becoming vital to the evolving Internet. As e-commerce and e-business become everyday realities, consumers want to control information about themselves, and at the same time to be able to share it with trusted parties. For their part, businesses want to serve consumers better by knowing more about them, while needing to comply with consumer privacy concerns and regulations. And from a technology standpoint, the rapidly growing Web services infrastructure requires a way to authenticate users that is not tied to network location or device. The solutions to these problems are all aspects of digital identity.

DIGITAL IDENTITY AND PRIVACY

Public awareness of privacy issues has been steadily increasing over the past decade, in large part because of the increasing use of the Internet and its ability to permit data sharing on a scale never possible before. Telemarketing, junk faxes, and unsolicited bulk e-mail ("spam") are small but annoying signs of the growing erosion of privacy. Consumers have begun to push back, not only by calling for legislation but by demanding products and services to protect their privacy. A good example is telephone services. Phone companies used to offer only one privacy service (an unlisted phone number). Now most offer Caller ID, which lets subscribers see who is calling before picking up the phone; Caller ID Blocking, which lets subscribers keep people who have Caller ID from finding out who they are; and finally Anonymous Call Blocking, which automatically rejects calls from people who use Caller ID Blocking.

A more serious problem is the relatively new crime of identity theft, in which an individual's personal information (usually stolen from a wallet or snatched from the mail) is used to impersonate him or her for the purposes of fraud. For example, an identity thief might obtain enough personal information to obtain credit and write bad checks in another person's name and then leave him or her stuck with the aftermath. Victims of identity theft often face years of frustrating harassment from creditors as they try to clear their records. Ultimately, every law-abiding citizen pays the price for these losses.

In the United States, civil libertarians have long been sounding the alarm about the amount of personal information stored in government databases. After September 11, 2001, many Americans have said they are willing to sacrifice some privacy to feel more secure, but to others, the existence of government surveillance systems such as Carnivore and Echelon points to an Orwellian future. Of what worth is security, they argue, if it comes at the cost of important Constitutional protections? Some also fear that safeguards against misuse of information stored in corporate databases are insufficient. They warn that employers, for example, might eventually use medical records to deny employment to those with unfavorable health conditions (or, as DNA testing becomes more precise, even to those with the *potential* to develop such conditions) by "partnering" with health insurers and hospitals to share this information.

These concerns have spurred a flurry of legislative action internationally. In the United States, the Gramm–Leach–Bliley Act (GLBA) makes financial institutions provide their customers with a written privacy policy, regulates how they can share their customers' personal data for marketing purposes, and requires them to allow consumers to opt out of marketing from nonaffiliated third parties. Also in the U.S., the Health Insurance Portability and Accountability Act (HIPAA) is intended to guarantee security and privacy of personally identifiable health information, while requiring that standard data formats be used in order to allow these data to move with the patient to competing healthcare and insurance providers. In the

U.K., the Data Protection Act requires firms to tell consumers upon request exactly what personal data about them the firms have in their files and also establishes guidelines for collecting, storing, and sharing this information. The European Data Directive establishes a clear and stable regulatory framework for the movement and storage of personal data for the entire European Union.

What impact will digital identity technology have on privacy issues? The long-term effects of any new technology are impossible to predict, but there are reasons to believe that many of these effects will be positive. Digital identity technologies allow individual users to set the terms under which their personal data can be shared with others and for what purposes these data may be used. Unless the requester of the data agrees to respect these terms, the protocol does not allow the data to be shared. In this scenario, digital identity provides a way for organizations to demonstrate compliance with privacy regulations. In fact, in the coming years, the authors believe that a key role of government will be to make agreements between digital identities legally binding, which has the potential to strengthen consumer privacy significantly.

The other, potentially more troubling, side of the digital identity coin is the technology's ability to link data stored in numerous databases into a coherent "virtual" view of an entity. Corporations and governments have collected thousands of pieces of information about every customer and citizen. Currently, this information is in thousands of separate databases, which makes it a limited threat to privacy. Digital identity has the potential to unify these databases not just within a single organization, but also across trust domains—meaning that the data kept about consumers by different companies might eventually be compared and coordinated. This sounds scary, but the technical and business challenges involved in this kind of consolidation of personal data are not trivial, and digital identity is hardly a "silver bullet" for solving them. Clearly, though, there is a vital role for government here, with legislation requiring companies to tell consumers upon request what personal data they keep, where they come from, and who else they will be shared with being a *minimum* requirement. An even stronger solution would be to require any data held by a corporation about an individual in a digital identity be electronically accessible and editable by the individual who is the source of the data. However, this might be achieved even without legislation if consumers demanded that the companies they dealt with simply subscribe to the data stored in their personal digital identity using permission-based links, rather than keeping their own private copy of the information. This is an area in which consumer lobbies and civil liberties watchdog groups could have significant leverage by educating the public on the issues—and now is the time to start.

The government, in addition to regulating digital identity, also will become a significant user of this new technology. In the United States, the Federal government alone maintains thousands of databases in its hundreds of agencies; each of the 50 states (plus territories) also has a government, and there are thousands of county and municipal governments below them, each with its own databases about the citizenry. Consolidating all these bits of data could save billions of taxpayer dollars and enable a new level of government effectiveness and service. In particular, the United States's new Homeland Security department is intensely interested in digital identity technology as a means of allowing the country's various local, state, and Federal intelligence, law enforcement, and emergency response agencies to aggregate the data each agency has, until now, maintained separately. Homeland Security has already stated their intent to aggregate information collected by other parties (such as library checkout records or bookstore purchase records) as well.

For obvious reasons, this new capability worries civil libertarians deeply. In itself, the deployment of digital identity technology for law enforcement and intelligence is not particular cause for alarm; with 280 million people living in the United States, it is unlikely that a given individual's information would come under scrutiny without good reason. The more important issue is defining what constitutes "good reason." Once the information has been aggregated, it can be "mined" to look for suspicious patterns, and the individuals identified could then have their activities tracked electronically. (Of course, data mining can be done without digital identity technology, and to some extent it already is.) Any surveillance technique based on data mining will inevitably result in a lot of innocent people being put under surveillance, at the very least having their records flagged for closer attention. It may be just a computer scan at first, but at some point this practice becomes an invasion of privacy and a violation of civil liberties.

To be fair, law enforcement is involved in an information arms race. There have already been reports in major news outlets of organized crime using data mining to identify and assassinate informants in their ranks, and it would be ridiculous to deny the same technology to police merely because it might be abused. But the need for security must be balanced with the need to preserve the civil liberties that modern democracies were founded upon. Citizens should demand due process and accountability from their governments and must remain vigilant to abuse. Data mining can and should be held to the same legal standards and privacy expectations as other forms of police work.

DIGITAL IDENTITY AND WEB SERVICES

The increasing prominence of Web services as a new way to achieve cross-application and cross-domain data integration is also highlighting the need for digital identity. Because Web services are almost as new as digital identity, it is worth taking a brief look at them here.

The term "Web services" has caused a lot of confusion among the uninitiated. Most Internet users have bought something at amazon.com, searched the Web using Google, or used their banks' Web sites to check their checking account balances. These are services available on the Web, but they are not "Web services" in the sense we're talking about here. In the enterprise, the term "Web services" applies in a very specific way to automated application-to-application communications—it is an extension of the similar use of the word "services" in a local computing environment.

To further explain the difference between the Web and Web services, we will compare the two directly. A Web site is designed to be used by human beings, and it includes site-specific formatting and graphics for, among other things, branding purposes. For example, Bank of America's Web site looks nothing like Citibank's, and neither looks anything like Bank One's. This is inconvenient if a computer program, rather than a person, wants to access data offered by a Web site. Suppose, for example, a user of a personal finance program wants to automatically import her checking register into the program from a bank's Web site. To do this, the personal finance program needs to know how to find that information on every bank's Web site. That would require the developers of personal finance programs to provide a different interface module (to extract the desired information from the Web page) for every bank in the world, and this is completely impractical. A much better answer is a standard interface for all bank Web sites that lets personal finance software, once properly authenticated, retrieve a customer's checking register.

The Web services paradigm is an attempt to address this problem and others like it, which are legion. Four key standards are involved. First, the Extensible Markup Language (XML) is an established W3C format for data representation (World Wide Web Consortium, n.d.). The Simple Object Access Protocol (SOAP) is a W3C specification that defines how programs running on one computer can request that certain operations be performed on another. The Web Services Description Language (WSDL) is an XML schema for describing a Web services interface in a standard, machine-readable format. Lastly, Universal Description, Discovery, and Integration (UDDI) allows a Web site to automatically discover which services are available from a given provider and determine how they should be accessed. All of these either are already open standards or are in the process of being standardized, and an overwhelming number of major industry players (including IBM, Microsoft, Sun Microsystems, HP, and BEA Systems) have thrown their weight behind the concept.

With Web services, applications and portable devices can access specific information and services over the Internet without a full-fledged Web browser. When a harried traveler is checking a flight departure time on his cellular phone, he or she doesn't want the whole American Airlines Web site, and he or she doesn't want to try to use Internet Explorer on the phone's tiny screen. He or she just wants to punch in the flight number and find out whether he or she can make the flight or not. This is a perfect application for Web services. At the same time, Web services allow Internet applications to begin communicating directly with each other, as well as with users. For example, a Web portal might use Web services to retrieve top headlines from a news site and then format them according to the portal's own look and feel before delivering them to the user.

As with Web sites, some Web services are public, whereas others are accessible by subscription only or only to certain users (for example, a company might operate some Web services only for its employees). It is necessary to authenticate users before allowing them to use these restricted services. Requiring separate credentials for each service quickly becomes a headache, especially when a single user might access the service from a mobile device, from one or more Web portals, and from a desktop application. What is needed is some way to ascertain the identity of users no matter how they access a given Web service, to model users as independent digital entities rather than only representing the devices they use. *This is digital identity.* To complete the circle, digital identity services can themselves be provided as one or more Web services, making them widely available to users everywhere.

DIGITAL IDENTITY ILLUSTRATED

A simple example will serve to illustrate the basic model of digital identity. In modeling an information system, a common object-oriented methodology involves thinking of the system in terms of actors who engage in transactions with each other. For example, when a customer withdraws money from an automated teller machine, there are three actors: the customer, the ATM, and the bank mainframe with which the ATM communicates to verify the customer's balance before dispensing money. The actors and the actions between them can be represented using a sequence diagram like that in Figure 1.

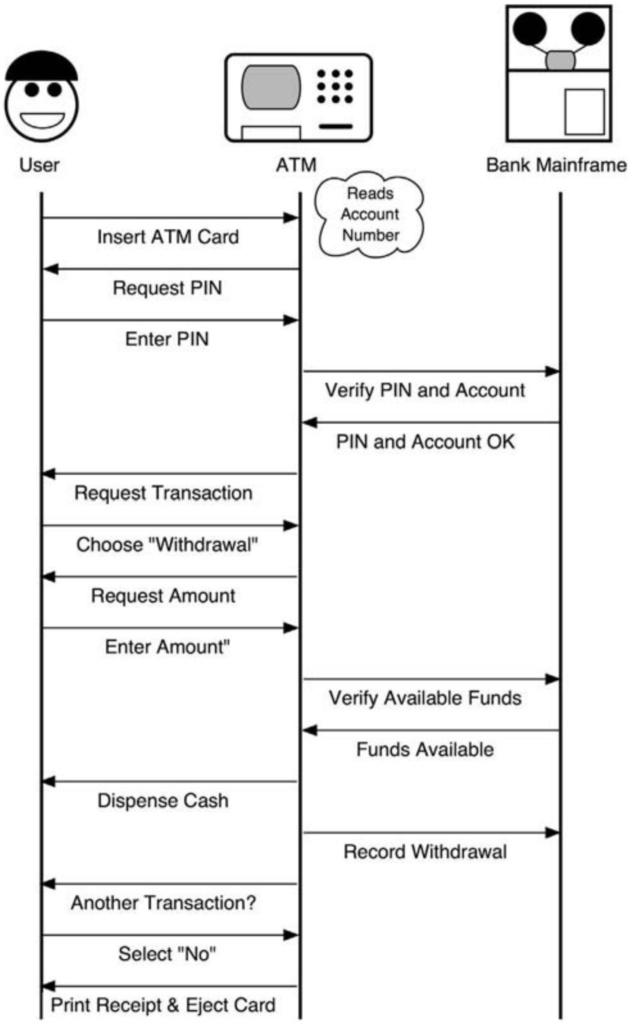


Figure 1: Sequence diagram for ATM withdrawal.

In this use case, the ATM has interactions with the user in which it asks for information such as a personal identification number (PIN) and the amount to be withdrawn. The ATM also has interactions with the bank mainframe, asking it to verify the account's current balance and telling it to record the withdrawal. (For clarity, this diagram does not show exceptions, such as what happens when there are not enough funds in an account to cover a withdrawal request.) These various interactions are possible because the bank owns all the data involved in the transaction. The bank balance is stored in its computer (along with a PIN), and although the account number is stored on the ATM card which the customer carries, that number is issued by, and thus in that sense belongs to, the bank.

The bank has plenty of other information on its customers, too: their mailing addresses for statements, their Social Security numbers (SSNs) for reporting interest income to the IRS, and their complete transaction histories. Though it describes customers, this information, too, belongs to the bank in the sense that customers do not have the authority to change it, even though they are the source of much of it. If a customer wants to change the address to which statements are mailed, for example, he or she must provide the new address to the bank so it can be entered into their systems.

The ATM system works because the all the actors involved in ATM transactions can trust each other, and they trust each other because the bank controls all the information involved. The ATM knows that it can trust the balance data, because it knows it is connected to the bank's mainframe and not to some other computer. The bank's mainframe knows that it can trust the ATM, because it has provided credentials that prove it is an ATM owned by the same bank. And the ATM and bank know that they can trust the customer at least modestly, because someone has inserted a card with an account number issued by the bank and entered the correct PIN number for that account. In digital identity terminology, the bank's ATM system comprises a single *trust domain.* All the information in the system is owned by a single entity, and all the components of the system are deemed trustworthy. The bank considers the other information customers have given it (such as mailing address and SSN) to be trustworthy as well, because it is in its systems and thus under its control.

To keep this scenario simple, it was assumed that the bank that owns the ATM is the bank where the customer has the account. In the real world, of course, banks join ATM networks; customers can use their cards at any ATM in a network that their banks are members of (although the banks may charge fees). To achieve this, the member banks have had to agree on a common protocol for exchanging data. Each bank has had to decide what credentials it will accept from other banks' ATMs to verify that they are in fact legitimate ATMs and what credentials the users of these ATMs will have to provide. They have also had to agree on what data will be provided over the network, and they have taken steps to protect the security and integrity of the data as they flow from point to point. In doing so, the banks have enlarged their trust domain to encompass all banks on the network, at least for ATM-related transactions.

With digital identity standards, trust domains can be expanded even further. What if not just banks but *every person and business in the world* could communicate and exchange data with anyone, just like the member banks of an ATM network? What is necessary to establish trust between all these entities? And what will such a world be like?

DIGITAL IDENTITY AND THE CONSUMER

Consumer digital identity applications are the obvious place to start exploring digital identity, as these have received the lion's share of media attention. Microsoft Passport, the Liberty Alliance, and Extensible Name Service (XNS) from the XNS Public Trust Organization (XNSORG) all initially led with consumer-facing applications.

The *digital wallet* is the most common example. A typical digital wallet contains a customer's mailing address, his or her preferred shipping addresses, and information about the credit cards he or she frequently uses for online purchases. When using a digital wallet to check out at an electronic storefront, the customer simply chooses the address and credit card to be used for the order from a pop-up list, rather than being forced to enter this information at each new store. Even more conveniently, when the customer moves or is issued a different credit card number, he she simply changes the appropriate fields in the wallet—once. The simpler a site makes the purchase process, the more likely it is to close the sale (Amazon's patented One-Click ordering is a case in point), so the digital wallet is potentially a big win for both consumers and merchants—as well as for banks, who earn a fee on credit card transactions.

Single sign-on (or SSO) is another consumer service. SSO lets users log on, or *authenticate,* at one Web site and then use that authentication at other participating Web sites. For example, after users log onto a Microsoft Passport account, they can get access to Hotmail, the Microsoft Network, Expedia, eBay, and other Web sites without having to enter their passwords again. It is particularly convenient to combine this with the digital wallet feature, making the wallet even more attractive for users.

Microsoft Passport is currently the most widely deployed single sign-on and digital wallet service at present, although adoption by consumers, vendors, and content providers has been slower than Microsoft had hoped. (A large percentage of the sites that accept Passport are owned by or affiliated with Microsoft.) Part of the reason for the slow uptake is that Passport is, at this writing, a centralized service. Microsoft stores all Passport data, including transaction records, on its own servers. Early technology adopters tend to be just the sorts of users who will insist on the privacy and security of their personal information, and Microsoft will have to provide more than convenience to win their trust. The Liberty Alliance specifications and XNS both support a feature called *federation,* where users can choose from a number of hosts for their personal data and cooperating sites can share authentication seamlessly, much like an ATM network.

Microsoft has stated that Passport will also move to this model.

But although digital wallets and SSO are part of digital identity, they are by no means the whole picture. It is, in fact, possible to implement these sorts of applications without even really modeling the underlying identity. A more robust conception of digital identity provides an infrastructure for a wide variety of applications.

THE IDENTITY WEB

To better understand the emerging vision of digital identity, we will look at the characteristics of a hypothetical full global identity infrastructure. Keep in mind that no such thing exists yet, and no one organization will build it. Instead, people and companies will begin using smaller identity-related applications as these applications become compelling. In fact, at first, users might not even realize they are using something called "digital identity." But as more and more people use identity services, they will naturally want these services to work together, further increasing their value in a classic network effect. This in turn makes additional identity applications more and more compelling, which brings more users into the identity fold, which further increases the network effect. This eventually results in the *Identity Web*—a global interlinked digital identity community, the virtual world in which, with the proper permissions, everyone can share anything with anyone.

To benefit from the network effect, naturally, identity applications must be interoperable, which means that a standard identity protocol must be adopted by all players. As of this writing there are two key digital identity standards emerging: XNS and the specifications from the Liberty Alliance Project. Both are open public standards. XNS is licensed by the not-for-profit XNS Public Trust Organization (XNSORG) (see XNS Public Trust Organization, n.d.), which recently contributed the XNS specifications to a new technical committee at OASIS (the Organization for the Advancement of Structured Information Standards), a global not-for-profit consortium that already maintains many of the key XML interoperability standards (Organization for the Advancement Information Standards, n.d.).

The Liberty Alliance is publishing its own specifications, though it may eventually contribute its work to another established standards body (some public reports have mentioned OASIS). Liberty's 65 members include American Express, MasterCard, Visa, America Online, Intuit, RSA Security, Sony, Sun Microsystems, Verisign, Fidelity Investments, Novell, Citigroup, Hewlett-Packard, Nokia, Vodafone, General Motors, Cisco Systems, Bank of America, PriceWaterhouseCoopers, EDS, Cisco Systems, Earthlink, OneName, and many others. This impressive array of members demonstrates the importance of interoperability to major corporations and provides enterprise customers with the assurance that Liberty standards will be viable from both technical and business standpoints (Liberty Alliance, n.d.).

The following sections draw on concepts from both of these emerging standards, with particular focus on XNS—largely because the authors are coauthors of the XNS specifications, but also because it reflects the broadest vision of digital identity available in the marketplace to date.

IDENTITY DOCUMENTS AND ADDRESSING

In XNS, the core concept of digital identity is the *identity document,* an XML document that stores the data for which any given real-world entity is authoritative. People own their names and addresses, for instance, so this information would best be stored in an individual's personal identity document. An identity document can represent a person, a corporation or other organization, a software application, a server, or even a directory category such as "plumbers"—any entity that has attributes that might be useful to share with others. (In the case of the "plumbers" category, the attributes being shared might be a definition of the category "plumber," a list of other associated categories, and a list of pointers to other identity documents that all belong to plumbers—a digital identity Yellow Pages.) Applications can query an identity document for just the attributes they are interested in; it is not necessary to retrieve and parse an entire document.

The identity document is hosted by an *identity service provider* on an *identity server.* As with Web hosting services, in the Identity Web there would be a multitude of identity service providers a user could choose to host his or her identity. Strong encryption would be used to ensure that only authorized users had access to the information in the identity document. Because digital identity is by its nature peer-to-peer (any identity can talk to any other identity), particularly savvy users could even choose to host their identities on their own computers for additional peace of mind.

To access the identity document itself and link to the information stored in it, an addressing scheme is needed. Standard URIs (uniform resource identifiers) such as HTTP URLs, which are used to address documents on the Web, do not have the full range of functionality required by digital identity infrastructure. For example, links between identities should continue to work even if a digital identity moves to another host. (The details of how links are used to share data will be revealed shortly, but for now it suffices to note that broken links would render some shared data inaccessible, limiting the utility of sharing data in the first place.) To provide this higher level of functionality, XNS uses a new form of URI called an *XRI* (extensible resource identifier). An XRI can be either an identity ID or an identity name. An *identity ID* is a string of numbers and other characters that forms a permanent address for an identity. Technically, an identity ID resolves to the URI where the given identity is hosted, and although the identity ID never changes, the URI it resolves to can change if the identity is moved to a different host. Each attribute stored in an identity document, as well as the identity document itself, has an ID, and this ID, too, is the same as long as the object exists. In XNS, document and attributed IDs are assigned and resolved by the ID service.

Identity IDs, however, are identifiers only a computer could love. They will be assigned more or less arbitrarily and are structural rather than semantic. Typically, a lot of numbers and punctuation will be involved, making IDs even less human-friendly than phone numbers. Just as DNS was invented to allow human-friendly host names to be assigned to computer-friendly IP addresses, so digital identity needs a way to make identity IDs easier for people to use. Enter the *identity name.* The XNS Name service allows any object with an identity ID (that is, an identity document or an attribute stored in it) to be assigned any number of names, which can be changed at the identity owner's whim. The naming service also allows an identity owner to organize the attributes stored in his or her identity document into categories or folders. For instance, phone numbers could be stored in a collection called "phone," allowing a cellular phone number to be named something like "=JohnSmith/phone/mobile." It is important to remember that although an identity document or an attribute might be located by name initially, when a link is established, it always uses the unbreakable ID.

In the beginning, it is likely that organizations will operate their own ID and name resolution services for customers and partners. For example, a bank might offer identity services to its customers, which would work with the bank's merchants but with no one else. In other words, the bank would provide ID and name resolution services only for its customers and merchants. The identities findable through a single resolution service are referred to as a *community.* Using the process of federation, which was mentioned earlier in the context of consumer identity services, communities can agree to share these services and thereby merge their smaller communities into larger ones. As identity services become more popular, a demand will eventually arise for *global community services* that, like the global Internet's DNS, allow any identity in the world to be found by its name or ID.

IDENTITY LINKING

Once attributes have been stored in identity documents that can find each other by ID or name, the stage is set for letting these identities share data. *Linking* is the way information from one identity is shared with another (see Figure 2). Links are almost always made using IDs, and because IDs never change over the life of an object, links between IDs never break while the linked objects exist. When one identity needs information that is controlled by another identity, the first identity simply links to the data in the second, forming a conduit through which data can flow. Then the two identities agree as to how the shared data will be kept current in the second identity—that is, how they will be synchronized.

An example will make the linking concept clearer. A person's mailing address, for all practical purposes, belongs to the person who lives at that address. (Street addresses are actually assigned by the city, but the city rarely changes them. More commonly, people change their own addresses—by moving. Thus, a mailing address can reasonably be treated as an attribute of an individual's identity.) A bank needs a mailing address to send statements, so the customer provides this information, and the bank puts it into its system. As noted earlier, however, the bank controls that copy of the address. If a customer moves, he or she must ask the bank to change it—even though it is *his or her* address! Change-of-address forms are so unremarkable that it barely seems that this is a problem worthy of solving. But just envision all the copies of information that people and companies keep about other people and companies; a vast amount of storage is used for redundant copies of all this information, and an equally vast amount of effort is expended on keeping all the copies current.

Linking addresses this problem head-on. In our example, the bank would simply link to a mailing address stored in the customer's identity document, using the address's immutable ID. To the bank's systems, it looks just as if a copy of the customer's mailing address is stored in their database. However, the customer's copy of the mailing address, stored in his or her identity document, is *authoritative,* meaning the customer can change it and the bank has to go along. When the customer moves, he or she simply

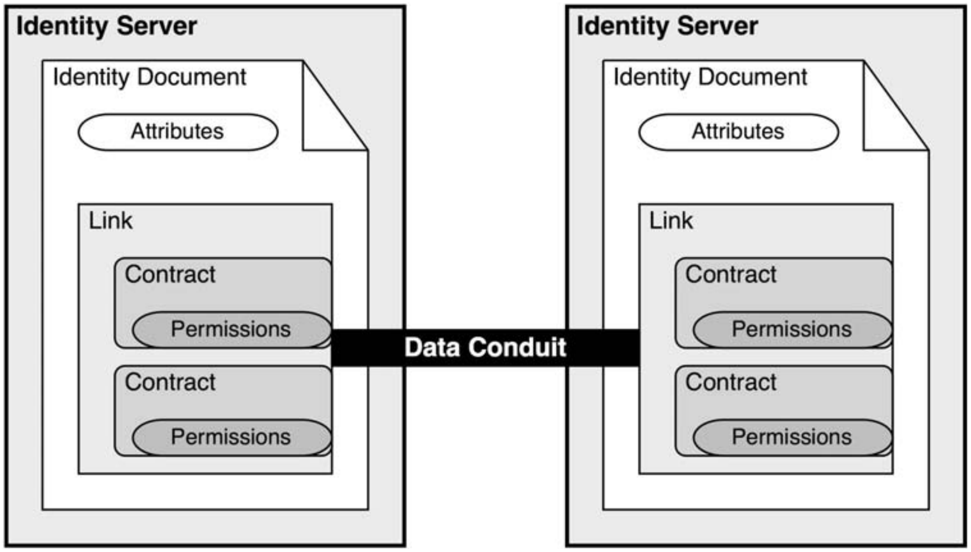


Figure 2: An identity link establishes a conduit between two identity documents.

updates the address stored in the identity document, and the identity service provider then sends a digitally signed copy of the updated address to the bank. The bank's system verifies the digital signature to make sure the update is genuine and then automatically stores the new address in the customer database. In essence, by eliminating redundancy, digital identity turns the world into an object-oriented, distributed, well-normalized database.

Of course, some of the technical details have been glossed over. The XNS protocol actually gives the bank in this scenario several options for synchronizing with a customer's identity data. For example, instead of having updates "pushed" to them, the bank could simply retrieve a copy from the customer's identity document any time they needed it by sending a digitally signed request to the customer's identity provider. Or the bank could just be notified that the address has changed and use some other method to update its database (for example, mailing a confirmation form to the new address to verify that it is correct). The exact details of how updates are handled will depend on the implementation, but regardless of how updates are accomplished, linking of identity data is arguably the most important concept in the sphere of digital identity.

Linking also provides the means to consolidate multiple digital identities that represent the same real-world entity. For example, a bank maintains a digital representation of each customer's account balance and activity. This information actually is part of an individual person's identity—it is his or her account, money, and transactions—but, for security reasons, the bank must maintain the authoritative copy. (It just wouldn't do to let people change their own account balances!) In an identity-centric world, the bank might do this by creating an identity document on its identity server for each of its customers. The attributes of these documents would be customers' various accounts, and each account would have a balance and a transaction history associated with it, among other things. So an individual's bank-hosted identity represents the real-world person just as much as his or her personal identity, hosted elsewhere, that contains his or her mailing address. We might call one the "financial identity" and the other the "household identity." By creating a link between the financial identity and the household identity, the customer can create a single *virtual* identity document that allows him or to view all his or her personal information (including the portions maintained by the customer, the bank, the customer's doctor, and other parties) in one place. Any number of third-party digital representations of a single real-world identity could be linked in this fashion.

A final application for linking is to intentionally split a single identity's attributes into two or more locations for security reasons, for much the same reason that banks maintain account balances rather than allowing customers to do so. For example, using linking, a cellular phone can become a digital wallet. By storing credit card numbers on a smart card inside a phone and configuring the phone to require its owner to "unlock" it using a PIN before it releases a credit card number, users can enjoy increased security for online shopping. To complete an online purchase, the phone must be turned on, it must have the user's (encrypted) smart card installed, and the user must approve the purchase by entering the PIN. This makes it very difficult for anyone but the legitimate owner to use a digital wallet for purchases without the owner's knowledge or approval. Using linking, however, users could still see this information as part of their main identity and update it as easily as any other information in their profiles—once they have logged in, of course.

DATA SHARING PERMISSIONS AND CONTRACTS

So far, this hypothetical identity Web is a framework for sharing everyone's identity data freely with everyone else. This utopia cannot exist, because people simply are not willing to share all of their personal data with everyone who might want it. Indeed, people are very discriminating with their personal information. We only give our credit card information to merchants we trust. We only give out direct e-mail addresses to others if we believe we won't get more junk mail by doing so. We pay extra to have our telephone numbers not listed in the telephone directory. Therefore, one of the driving principles of digital identity is to give control of private data back to the people and organizations to whom it belongs. A digital identity infrastructure must support flexible privacy control at the protocol level to gain the trust of users.

The XNS digital identity infrastructure handles privacy through a data structure called a *contract* that defines the terms under which data may be shared. For example, a contract between a personal identity and a merchant identity might specify that the shipping address stored in the identity document will be used only to fulfill the current order and not to mail advertising; the merchant will not retain the address in its systems any longer than is necessary for the purposes of shipping the order; and the merchant may not share the address with any other parties without customer consent.

How are contracts established? By a process called *negotiation.* (The XNS Negotiation service is an implementation of the privacy framework developed by the International Security, Trust, and Privacy Alliance, ISTPA [International Security, Trust, and Privacy Alliance, n.d.]) In a transaction between a customer and a merchant, the merchant identity would present to the customer identity a proposed contract indicating what information is needed, what it will be used for, how long it will be kept, and with whom it may be shared. If the customer has previously told her identity service provider what terms he or she finds acceptable and provided preferred values for the kinds of data the merchant wants, negotiation could be completed entirely without human intervention. More commonly, the customer would insist on having the opportunity to review and approve the terms and select the data to be shared before finalizing the transaction. The resulting contract records the mutually agreed upon terms and governs the links between the two identity documents between which the data will be shared. When a permission requirement has been established on a data attribute stored in an identity document, it becomes impossible for other identities to access that data without first agreeing to a contract.

Because the resulting data structure is *called* a contract, the obvious next question is whether it is legally binding. At this writing, the issue has never been tested in court. But the law in many countries is moving steadily toward accepting digital signatures as legal proof of assent. Using established cryptographic techniques, it is possible to provide assurance of the identities of the parties agreeing to the contract and to prevent repudiation, so eventually contracts between digital identities may be as binding as written contracts between real-world parties.

Even without the weight of the law behind them, data-sharing contracts will make it easier for consumers to access and update the personal information that businesses keep about them, as well as allowing consumers to hold businesses more accountable for the ways they use their customers' data. Although some companies might not care for this level of accountability at first, eventually businesses will embrace it wholeheartedly, for a simple reason: contract-based data sharing provides a foundation of trust for building closer, more profitable relationships with customers, employees, suppliers, and partners.

OTHER DIGITAL IDENTITY SERVICES

Being able to locate identity documents and forge permission-protected data-sharing links between them makes a wide variety of digital identity applications possible. This is the foundation of XNS and also of Liberty Phase 2. Indeed, analysts predict that data sharing will be the single largest purpose for which digital identity technology will be used, especially in the near term. Beyond sharing, however, digital identity enables other services that will make the Internet and many common communication functions safer, easier, and faster.

Many of these are *trust services*—services that make it easier to perform trusted digital transactions, just as credit cards made it easier to perform trusted mail-order transaction. For example, an *authentication* service is needed to prove that a digital identity represents a real-world principal and to assert that real-world identity across trust domains. Once a user logs into his or her digital identity, in other words, other services can tell that he or she is the same person who created the identity, and therefore these services can be assured that they are dealing with a single individual at all times. This enables single sign-on and customization features and also supports the creation of contracts between identities.

Certification service allows digital identities to make assertions about the truth of the data stored in them. A third party, called a *certifying authority,* provides a cryptographic signature that essentially claims that it has inspected the data in question and found them to be complete and accurate. A real-world example is a driver's license. It would be much better to keep a digitally signed version of a driver's license in an identity document rather than just including the raw data the license contains; the digital signature would allow this credential to be used as proof of age, for instance. Without a verifiable digital certification from the state of origin, the merchant would have no reason to trust the information. Certification supports the mapping of real-world identity credentials into the digital realm and thus gives digital contracts their force.

A *session* service combines authentication and certification to support a browser-based single sign-on (SSO) solution. The third party in this case is a *session server,* which issues a cryptographic token certifying that acceptable credentials were presented to begin a session. This *session token* then either is accepted by various Web resources in place of traditional credentials (such as a user ID and password) or is transparently mapped to acceptable local credentials. The Liberty Alliance version 1.0 protocols are essentially a federated session service.

A *hosting* service allows one identity to host another. When an identity document is hosted by an identity server, the hosting service is the service responsible for establishing the new identity's persistent address. In XNS, the identity server itself is represented by an identity document. Identities hosted by a server are registered with the server's *host identity* so that they can be located by name and ID.

Finally, in an extensible identity protocol, some way must be provided for applications to discover new identity services they do not know about. In XNS, this is accomplished with the Discovery service, which designates a *publishing identity* for each service's definition and allows the service's data formats and message descriptions to be retrieved in standard formats including XML schema (XSD) and Web services definition languages (WSDL). The Discovery service allows anyone to define new identity services that are completely interoperable with existing identity applications.

In XNS 1.0, the 10 services ID, Name, Discovery, Hosting, Data, Folder, Authentication, Certification, Session, and Negotiation—plus the Core service, which defines data formats and an abstract base class for messages—compose the base services. These are, in the authors' estimation, the absolute minimum requirements for a fully functional identity infrastructure. In a future version of the XNS specifications, additional services are planned, including the following:

Reputation service will allow the trustworthiness of an unknown identity to be evaluated based on the opinions of other identities;

Introduction service allows two identities that currently link to a third identity to cut out the middleman and link directly—essentially a "three-way negotiation" or "friend-of-a-friend" service;

Directory service provides the ability to register and locate participating identities not just by name and ID but by certain attributes (for example, locating all identities at a specific company or in a particular geographic area).

Thanks to the extensible architecture of XNS and its Discovery service, further services can be developed as needed by anyone who needs them.

DIGITAL IDENTITY AND ENTERPRISE APPLICATION INTEGRATION

Large companies contain vast repositories of identity data. An LDAP directory server contains identity data; so does a customer relationship management system, an accounting system, or an order processing system. Even a network management tool, an e-mail server, or a collaboration server could be viewed as containing identity data. (Anything that stores and manages data about entities can be considered as an identity application in a loose sense.) None of these tools are designed to talk to each other, even though the benefits of doing so might be enormous.

The problem is compounded in obvious ways when companies join forces; it is not uncommon for a large enterprise to have several separate, incompatible information systems performing the same functions for its various divisions, like a fossil record of the company's merger history. This is the redundancy problem from this chapter's very first example—the same information stored in many places, wasting vast amounts of storage and requiring enormous effort to keep in sync—written *very* large. The biggest corporations have hundreds or even thousands of different mission-critical databases stored in mutually incompatible systems around the world.

Addressing these issues in a comprehensive way all at once can be mind-bogglingly expensive and time-consuming. Designing, implementing, and testing just data format conversions can take thousands of engineer-hours. Then there is the expense of replacing the legacy systems with more modern software and the inevitable headaches that come with trying to tie it all together, not to mention the additional time required for retraining employees on the new systems.

Envisioning digital identity as an integration layer on top of existing enterprise information systems makes it possible to undertake enterprise application integration (EAI) in small, easily managed pieces. The legacy systems are kept intact and remain authoritative, with the data needed for a given application transparently mapped into identity documents using standard EAI *adapters* that serve as bridges between the identity server and the legacy systems. Using this approach, new integration applications can be built on top of the identity layer rather than talking directly to the enterprise layer and can be developed with modern object-oriented methodologies, with the legacy systems treated more or less as black boxes. Companies can begin with the proverbial "low-hanging fruit"—the most obvious enterprise integration projects with the clearest potential for substantial return on investment or the best chance to achieve a new competitive edge. As these projects prove themselves out, more projects can be undertaken with increasing confidence, until at some point the network effect comes into play and the ROI of identity solutions already implemented increases exponentially by allowing new connections to be made. In the coming identity-centric world, it will eventually become a competitive disadvantage to store identity-related data in places where they cannot be shared (given appropriate permissions) with others.

An example will serve to illustrate how digital identity can be used for integration. Suppose two mail order companies merge into one bigger company that now has two separate systems for processing orders and for mailing catalogs. An identity-based approach to integrating these systems, to allow customer service representatives to access all data stored in both systems without even having to know that there *are* two separate systems, would be to create an identity document to represent each customer. By means of adapters, individual data elements in identity documents would be retrieved from the existing systems. The identity document stores any record keys necessary for finding the appropriate data in the two systems. When an application asks the identity server for the catalog mailing address from a customer's identity document, the identity server retrieves the data "on the fly" from the mailing list application. Requesting the last order date causes the identity server to query the order-processing system for this bit of information. If the customer's entire identity document were requested, all the information needed from all systems involved would be retrieved behind the scenes, by means of adapters, and stitched together into an identity document conforming to the syntax required by the standard (for example, XML). When an application changes something in the identity document, the adapters write the changes back out to the legacy systems. Adapters can also take care of any necessary protocol details (e.g., interfacing with the desired network stack), do encoding conversions (e.g., EBCDIC to Unicode), and provide caching services and even encryption and decryption.

The application never needs to know or care that any of this is happening; it uses the same simple, standardized object-oriented methods to read or write data regardless of where it is stored and how it must be accessed. Existing applications that are used with the legacy data systems can remain authoritative or can even be run simultaneously with the identity-based applications. In fact, if desired, the legacy systems can even be switched out from under the identity layer—suppose they need to be upgraded for other reasons—and, assuming the appropriate adapters have been written for the new systems, the identity-based applications will never notice. The adapter approach also makes it possible for system integrators to build generic solutions for common integration problems; for each installation, they need only write adapters and perform the specific customization necessary to get the new applications running on top of a particular client's legacy data stores. Figure 3 shows how the resulting system looks at a high level.

Once a data source becomes identity-enabled, it is easy to add functionality incrementally. The first step might be to integrate the order processing and mailing list systems of two merged companies, as described above, so that customer data are more accurate and customer service staff can work more efficiently. The next step might be to add a Web interface to the system so that customers can access and update this information on their own, reducing staffing needs for the call center. Next, the accounting system might be integrated to allow customers to see and pay their invoices online. Then the company might offer to integrate its order-processing system with its identity-savvy institutional customers' inventory systems, so that their stock can automatically be replenished when it runs low. Each step adds real value, each has measurable ROI, and each can be accomplished in a relatively short time with relatively few development resources thanks to the standardized identity programming interface.

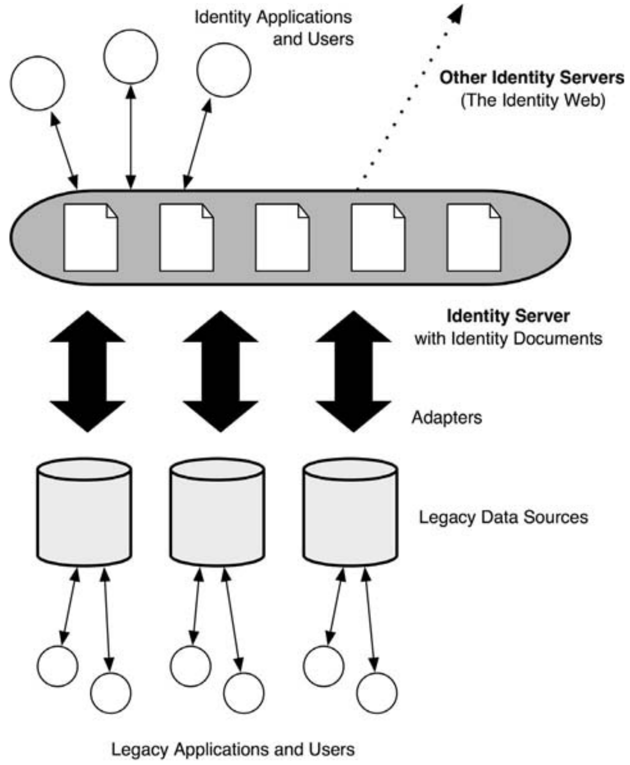


Figure 3: Digital identity as an enterprise integration layer.

Stunning feats of integration like these can be accomplished without digital identity—Wal-Mart's integration of its suppliers into its own data processing operations is the textbook example. However, doing it the old-fashioned way requires expenditure of money and development resources in direct proportion to the number of data already stored in legacy "silos," and it is difficult for most enterprises to justify the expense of many such projects on a ROI basis. Identity-oriented development promises to do for enterprise systems integration what object-oriented programming did for software engineering: make it simpler, quicker, less error-prone, and less resource-intensive. Suddenly, many projects that did not originally make financial sense look feasible, and they look even more attractive as the network effect made possible by open standards takes hold.

Our prediction is that the enterprise is where most identity-related development will occur over the next few years. Consumers will come to rely on identity-based services provided by these companies more and more, often without knowing (or needing to know) that these services are in fact founded on digital identity technologies. For example, companies will start allowing customers to share their account information with their strategic partners. Meanwhile, grass-roots understanding of privacy issues will slowly but inexorably lead to consumer demand for identity applications and for true control over their own data. Consumer-protection legislation and regulations will drive enterprises to adopt technologies with strong privacy safeguards, lending additional weight to the benefits of digital identity. Network effects will, finally, lead to the Identity Web, a global hyperlinked collection of personal data shared, by mutual agreement, among people, companies, and other organizations. By the end of this decade, the most compelling reason to use identity-based solutions in the enterprise will be that your customers, vendors, partners, and competitors are already using them.

THE IDENTITY PLAYERS

This chapter has already touched on the efforts of One-Name Corporation, XNSORG, and the Liberty Alliance in the realm of digital identity and has touched on Microsoft's Passport. We can characterize these efforts as "pure" identity plays.

Liberty's Phase I specification defines a single sign-on (SSO) service based on the security assertion markup language (SAML), wherein existing accounts at multiple sites that have agreed to share authentication can be linked. Sun Microsystems is already supporting Liberty Phase 1 in its directory server. Phase 2, due early in 2003, is a more sophisticated protocol that allows any Liberty-compliant service to federate with any other—without pre-existing partnerships—and to share more attributes between accounts.

Unfortunately, Liberty proceedings are not open to the public, and until the specification is officially released, Liberty members are subject to strict nondisclosure agreements. For this reason, few details about Liberty's plans for Phase 2 are available to the general public. However, it is worth noting that the consortium's strategy has generally been to use existing standards when possible, as evidenced by the use of SAML in Phase 1; it is therefore not unreasonable to expect Phase 2 to be built at least partly on applicable Web services standards.

Microsoft's trustworthy computing initiative, popularly known by its code name, "Palladium," aims to add permissions to document-sharing of all kinds through a combination of security and identity technologies. The industry perceives trustworthy computing as being driven in large part by the desire of content providers (such as the music and motion picture industries) to control how their content is being used and protect it from piracy, but Microsoft recognizes that consumers will use a system with such restrictions unless it also provides compelling benefits for them as well. The carrot Microsoft offers with Palladium is to give *everyone* the ability to apply the same industrial-grade restrictions to information he or she creates. This in itself is not strictly an identity-related development; however, also as part of Palladium, Microsoft is providing wider access to the Passport technology including, in an unprecedented move for Redmond, sharing Passport Manager source code with selected partners and customers. (Passport Manager is the software module that allows a Web site to join the Passport network and access shared user data.) As mentioned earlier, Microsoft has promised that Passport will support federation based on the Kerberos 5.0 standard soon, so individual communities can begin using it now and decide later how they want to hook up (Microsoft Trustworthy Computing and Passport, n.d.).

These three are not the only players, though—only the most visible. Enterprise software vendors are also moving into digital identity. Many of the identity data in the enterprise are housed in directory servers, which are used to control access to corporate network resources as well as to store contact information. Manufacturers of directory and access-management products have a natural interest, therefore, in providing their customers with the integration and data-sharing tools they will need in an identity-centric world. Not surprisingly, vendors of access-control products—including IBM, Microsoft, Netegrity, Novell, Oblix, and Sun—are already beginning to tout digital identity functionality and interoperability for their products.

Current directory products do not scale well across trust domains—meaning that they are better suited to controlling access within an enterprise than across multiple enterprises—but some vendors already offer *metadirectory* products that can consolidate enterprise directories into one virtual directory. This monolithic approach does not offer the advantages of the Web architecture that will eventually evolve with digital identity; however, enterprises will continue to turn to these working solutions in the short term while identity standards shake out. At that point, it would be natural for these vendors to begin adding support for standards-based distributed identity.

As described earlier, digital identity has the potential to introduce a major paradigm shift in the EAI category. Where the traditional EAI approach is monolithic and expensive, the digital identity approach offers low up-front cost and incremental deployment. These are such powerful advantages that it is likely that EAI vendors such as BEA, SeeBeyond, TIBCO, and webMethods will eventually be compelled to use this approach and may in time become leading vendors of enterprise digital identity solutions.

THE FUTURE OF DIGITAL IDENTITY

This chapter has presented the authors' vision of how digital identity will, through links between identities that share data, eventually form an Identity Web similar to the current World Wide Web of documents. The initial growth of the Identity Web will be driven, like the early stages of the WWW, by early adopters, though early on, more businesses than consumers will see and exploit the potential of digital identity.

As digital identity infrastructure grows, it will have progressively wider applications in many product categories. For example, e-mail vendors could use digital identity protocols to provide permission-based mail filtering. Senders of mail would have to prove their identity and agree to a privacy contract before being permitted to send mail to a given address, potentially putting an end to "spam" once and for all. By the end of the decade, any product category that involves sharing information is likely to include digital identity features, particularly if it involves digital rights management (DRM).

Despite the possible privacy and civil liberties pitfalls, which continue to require close scrutiny, the concepts embodied by digital identity have the potential to have enormous net positive social and technological impact in the coming years. It is too bad that the word "revolutionary" has been tarnished by its application to so many half-baked technologies—if ever a new technology deserved that adjective, it is digital identity.

GLOSSARY

Adapter In enterprise application integration, a software module that serves as a bridge between an identity server and legacy systems.

Authentication The process of proving that a user is an authorized user of a network resource by presenting credentials such as a user ID and password (colloquially known as "logging on").

Community The group of identities that can be located through a given address resolution service.

Contract A list of mutually agreed permissions for sharing data between identities via an identity link.

Digital identity An emerging technology and application category that revolves around the logical modeling of real-world actors (such as people and companies) and their attributes and the permission-based sharing of these attributes between actors on a network.

Digital wallet A popular consumer digital identity application that stores customers' shipping addresses and payment methods to speed checkout at e-commerce Web sites.

Extensible name service (XNS) An open, XML-based digital identity protocol licensed by the not-for-profit XNS Public Trust Organization (XNSORG).

Extensible resource identifier (XRI) A new form of URI that provides the functionality necessary for addressing identity documents; can be either an ID or a name.

Federation The process whereby separate communities share address resolution services, thus merging two or more communities into a larger one and giving users a choice of which provider to use.

Global community services A publicly accessible identity address resolution service that allows any digital identity to be located by any other, much as the existing domain name service (DNS) allows any Internet-connected system to locate any other.

Host identity The identity that represents an identity service and is responsible for hosting other identities.

Identity document A document, stored on an identity server, that stores the attributes of some real-world actor (such as a person, a company, a device, or an application).

Identity ID An immutable, machine-readable identifier that resolves to the network endpoint where an identity document is stored; identity IDs never change as long as the document exists and are the foundation of identity links.

Identity link A "pipe" between two identities for the two-way synchronization of shared attributes.

Identity name A human-readable semantic identifier for an identity document that can be resolved to an identity ID.

Identity server Software that stores identity documents and provides identity services for those documents.

Identity service provider An organization that provides storage for identity documents.

Identity Web The global interlinked digital identity community (see *Community*), analogous to the World Wide Web of documents.

Liberty Alliance An industry consortium founded by Sun Microsystems to deliver and support federated network identity solution for the Internet.

Negotiation The process of agreeing upon permissions for sharing data between identities via an identity link. These permissions are embedded in a *contract*.

Passport Microsoft's consumer digital identity service, based on the Kerberos authentication protocol.

Publishing identity The identity that makes available the definition of an identity service so that the service can be discovered by identity-savvy applications.

Single sign-on (SSO) A digital identity application that allows users to log in once to gain access to multiple participating Internet sites, rather than having to enter credentials separately at each site.

Trust domain Technically, a set of network resources the access to which is controlled by a given directory server; in practice, the network resources a user can access with a single set of credentials; more generally, a set of network resources that trust each other, such as an ATM network.

Trust services Identity services that make it easier to perform trusted digital transactions; examples include authentication and certification.

Web services A new model for inter-application communication over the Internet using standardized protocols (XML, SOAP, WSDL, UDDI); the extension of the concept of local computing services to the Web.

CROSS REFERENCES

See *Authentication; Biometric Authentication; Digital Signatures and Electronic Signatures; Extensible Markup Language (XML); Online Communities; Passwords; Privacy Law; Web Services*.

REFERENCES

International Security, Trust, and Privacy Alliance (ISTPA) Web site (n.d.). Retrieved May 14, 2003, from http://www.istpa.org/

Liberty Alliance Web site (n.d.). Retrieved May 14, 2003, from http://www.projectliberty.org/

Microsoft Trustworthy Computing and Passport Web site (n.d.). Retrieved May 14, 2003, from http://www.passport.net/Consumer/default.asp?lc=1033

Organization for the Advancement of Structured Information Standards (OASIS) Web site (n.d.). Retrieved May 14, 2003, from http://www.oasis-open.org/

World Wide Web Consortium (W3C) Web site (n.d.). Retrieved May 14, 2003, from http://www.w3c.org/

XNS Public Trust Organization Web site (n.d.). Retrieved May 14, 2003, from http://www.xns.org/

FURTHER READING

Digital Identity World Web site: http://www.digitalidworld.com/ (Digital identity news site and "hub of the digital identity industry"; also organizes the Digital Identity World conference)

Electronic Privacy Information Center (EPIC) Web site: http://www.epic.org/ (Public interest research center in Washington, DC focusing on emerging civil liberties and privacy issues)

Electronic Frontier Foundation (EFF) Web site: http://www.eff.org/ (Grass-roots organization focused on protecting civil liberties at the interface where law and technology meet, including privacy issues)

Microsoft Security Web site: Privacy http://www.microsoft.com/security/ (Information about Microsoft's consumer identity, security, and privacy initiatives, including trustworthy computing, also known as Palladium)

U.S. Federal Trade Commission Privacy Initiatives Web site: http://www.ftc.gov/privacy/index.html (Official U.S. government Web site on privacy and digital identity issues, including identity theft)

Digital Libraries

Cavan McCarthy, *Louisiana State University*

INTRODUCTION

Defining Digital Libraries

Digital libraries combine the advantages of digital access with the services and quality information traditionally found in libraries. They can be characterized as the "high end" of the Internet: that sector of the Internet that maintains a tradition of shared resources and easy access while offering quality content from authoritative sources, such as the Library of Congress and major universities.

The author has developed the following definition of a digital library: a system that permits, via the Internet and the World Wide Web (WWW), easy access to a collection of high-value, quality digital content, which has been selected and organized to facilitate use and is supported by appropriate services. The digital content may reflect the traditional textual orientation of many libraries or take advantage of the WWW's facilities to deliver graphics and multimedia. The term digital library is also used to refer to quality information and referral services, whose organization reflects the traditional structure of library services.

The major textbooks for the field offer the following definitions of digital libraries:

> Digital libraries are organized collections of digital information. They combine the structuring and gathering of information, which libraries and archives have always done, with the digital representation that computers have made possible. . . . [T]he digital library must have content; it can either be new material, prepared digitally, or old material, converted to digital form. (Lesk, 1997, pp. xix, 2)

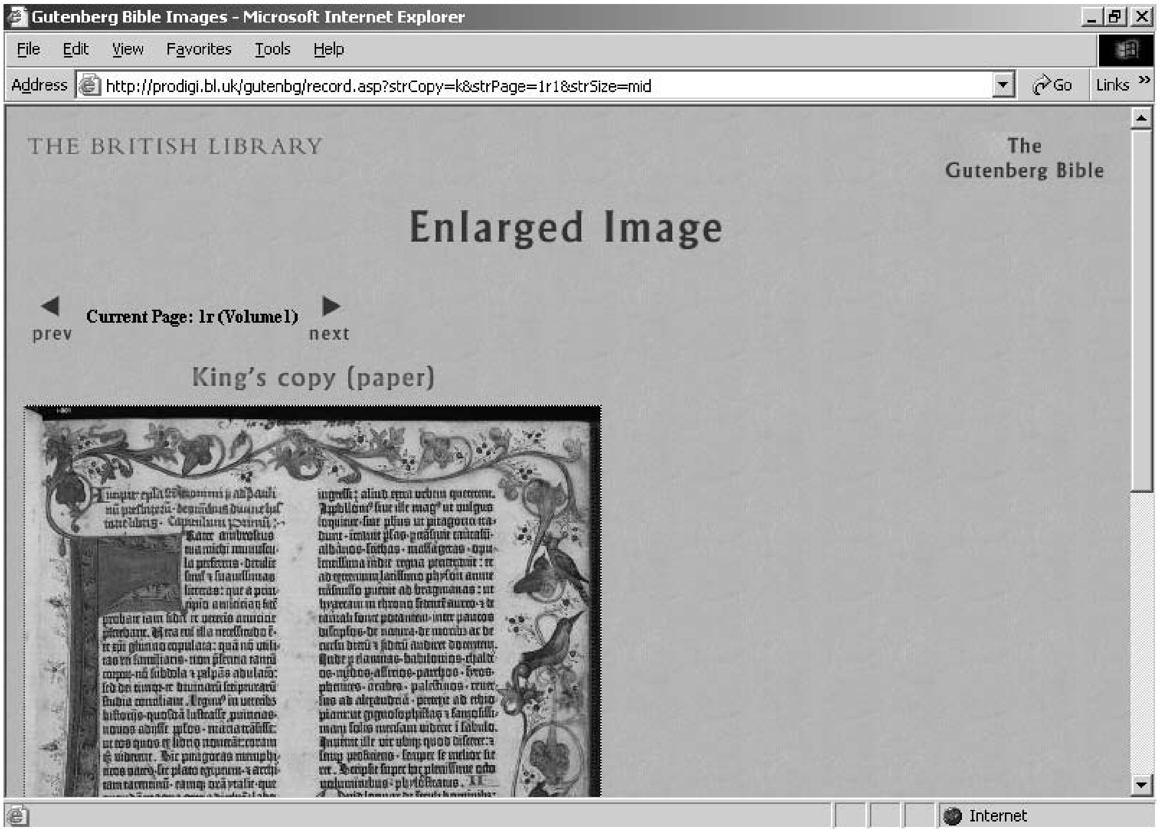
> An informal definition of a digital library is a managed collection of information, with associated services, where the information is stored in digital formats and accessible over a network. A crucial part of this definition is that the information is managed. . . . Digital libraries contain diverse collections of information for use by many different users. Digital libraries range in size from tiny to huge. They can use any type of computing equipment and any suitable software. The unifying theme is that information is organized on computers and available over a network, with procedures to select the materials in the collections, to organize it, to make it available to users, and to archive it. (Arms, 2000, p. 2)

As is common in areas that are still establishing their fields of activity, there are problems in reaching a precise definition. Three parallel terms exist: digital, electronic, and virtual libraries. These are normally used interchangeably (Saffady, 1995). In practice, digital library is the general term in North America whereas electronic library is preferred in Europe. Schwartz (2000) analyzed no fewer than 64 different formal and informal definitions of "digital library."

Advantages of Digital Libraries

The advantages of digital libraries are as follows.

Remote Access: It is no longer necessary to go to a library; resources and information services can be consulted from any location worldwide via the Internet without having to travel.



Gutenberg Bible Images - Microsoft Internet Explorer

Address http://prodigi.bl.uk/gutenbg/record.asp?strCopy=k&strPage=1r1&strSize=mid

THE BRITISH LIBRARY

The Gutenberg Bible

Enlarged Image

Current Page: 1r (Volume1)

King's copy (paper)

Figure 1: Decorated page from a Gutenberg Bible, digitized by the British Library.

Simultaneous Access: An unlimited number of persons can access the same resource, even read the same page of the same text, simultaneously.

24/7 Access: Digital libraries, like the Internet, are normally available 24 hr a day, every day of the year.

Digital Texts: These can be easily updated, printed, copied, and even manipulated. Digital texts offer exceptional facilities for manipulation or modification, but this must, of course, be done while respecting copyright and avoiding plagiarism.

Free: Access to many systems is free to users, reflecting the free access principles established both in libraries, especially public libraries, and on the Internet itself.

Indexing: Texts in digital libraries are frequently indexed word for word, and indexes often cover the contents of many texts. Go to the Victorian Women Writers Project, at http://www.indiana.edu/~letrs/vwwp/index.html, and enter a relevant word, such as "suffrage" to test the index. Traditional book indexing is typically limited to a few words per page, printed in the back of that specific book.

Browser Access: Popular browser software, such as Internet Explorer and Netscape, which Internet users are accustomed to using to obtain information, are used to provide access to digital libraries.

Rare Materials: Older, out-of-copyright, and unusual materials can be made widely available via digital libraries. Copies of the earliest printed books, Gutenberg bibles, are available, from the British Library (Figure 1; http://prodigi.bl.uk/gutenbg/default.asp) and from Goettingen, Germany (http://www.gutenbergdigital.de/). The Library of Congress' copy is being digitized as this article is being written.

Preservation of Originals: Old, fragile materials can be repeatedly examined through the Internet, with no possibility of harming the originals. For example, it is difficult to imagine a writing media more fragile than papyrus, but a large collection of papyrus documents can be examined at the Advanced Papyrological Information System (http://sunsite.berkeley.edu/APIS/) with no risk whatsoever to the originals.

Links: Digital resources can include hyperlinks to other relevant sources, such as contents lists, citations, other documents, information sources, specialized data bases, dictionaries, and human help (via e-mail or chat-style virtual reference service systems). A sophisticated example is offered by the Perseus classical studies system (http://www.perseus.tufts.edu/). This links each Greek or Latin word in the digitized texts to dictionaries and word frequency lists.

Reliability: Digital libraries do not have the problems common in traditional libraries, such as books being unavailable because they're on loan, misplaced, or lost; pages torn out or effaced; missing periodical issues, supplements, or indexes. Plus, being accessed via the Internet, this medium is always "up" or "on."

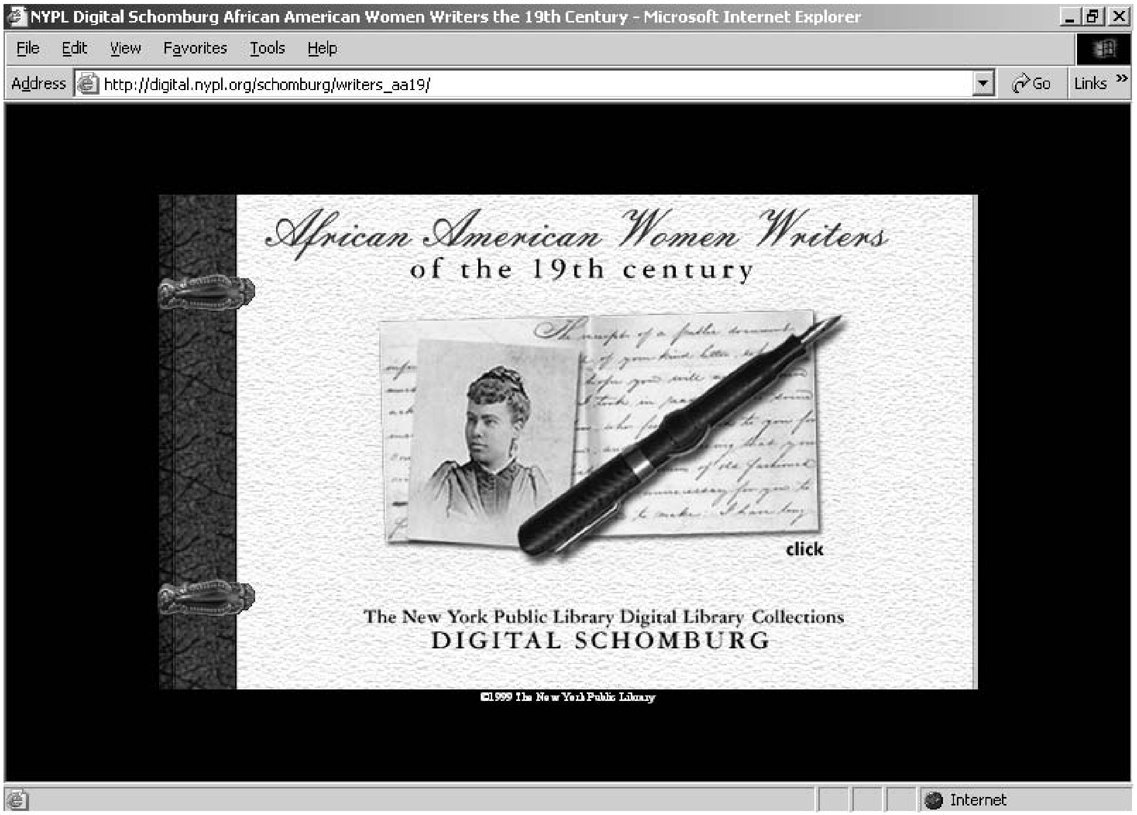

NYPL Digital Schomburg African American Women Writers the 19th Century - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://digital.nypl.org/schomburg/writers_aa19/
Go
Links
African American Women Writers
of the 19th century
click
The New York Public Library Digital Library Collections
DIGITAL SCHOMBURG
©1999 The New York Public Library
Internet

Figure 2: Texts by African American women writers of the 19th century have been digitized by the New York Public Library, Schomburg Collection.

Costs: Initial digitization and setup is expensive, and systems require constant maintenance and updating. However, the cost per use of heavily used systems can be low, and often materials are only available with great difficulty in traditional form.

Service and Social Impact: Digital collections can disseminate the culture and achievements of minority groups. A typical example is the texts by African American Woman Writers of the 19th century from the New York Public Library's Schomburg Collection (Figure 2; http://digital.nypl.org/schomburg/writers_aa19/).

Limitations of Digital Libraries

The limitations of digital libraries are as follows.

Reading from Computer Screens: People do not like reading lengthy texts on computer screens. Print quality is considered poor and computer screens have a horizontal, cinema orientation, rather than a vertical, page-format, layout. This leads to two subsidiary problems. (a) Printers are almost essential as adjuncts to digital library usage. Readers still prefer hard copies in many cases, and paper consumption continues to increase in the midst of the digital revolution (Emmott, 1998). (b) There is a lack of adequate specialized reading devices for digital texts, especially electronic books. Consumers have shown little interest in the early reading devices, which were considered expensive and more awkward to use than the printed books they were supposed to replace. Improved and cheaper laptop computers may fill the gap, but equipment has not quite reached that stage.

Limited Collections: At the moment, digital libraries exclude most modern commercially published books and most copyright-protected resources (i.e., exactly the texts for which there would be the greatest demand). The Internet Public Library links to "over 20,000" online texts (http://www.ipl.org/reading/books/). This is similar to the total cited by the Online Books Page, a wide-ranging catalog of online books, which covers about 16,000 titles (http://digital.library.upenn.edu/books/). Neither the Library of Congress nor any other major library intends to digitize the whole of its book collection. It is not certain when commercial publishers, who are afraid of losing control over digital versions of their texts, will make major investments in this field. Paid-access, subscription-oriented digital libraries also offer limited collections. Questia offers 45,000 complete books and 25,000 journal articles (http://www.questia.com); NetLibrary offers 40,000 books (http://www.netlibrary.com/).

Digital Divide: Access to high technology and hands-on training are essential for digital library consultation. This greatly limits their value in developing countries and communities that are disadvantaged, or unfamiliar with computers. If these difficulties can be overcome, digital libraries have a major potential for reducing divisions within society, because they make information more widely available.

Finance: Digital libraries require significant funding at a time when financing for libraries is stagnant. Libraries have paid significant sums to install computers and guarantee Internet access; many public and school libraries also pay to filter Internet access. Libraries now offer Web access to their catalogs and have subscribed to electronic journals, databases, and closed-access e-book systems. There is little extra money left to digitize their collections.

Personnel Problems: Librarians and systems analysts with training and experience in planning and mounting a digital library are still few in number. Education for digital libraries is based on "on the job" experience. The domain is not yet well developed within most academic curricula.

Content Description: It is often difficult to know what might be offered by a digital library before accessing it. Cataloging is complex; digital library materials can be available in different formats: ASCII/HTML/.PDF; indexed/not indexed; with high/low resolution images; and so forth.

CATEGORIES OF DIGITAL LIBRARIES

Despite their short history, digital libraries already demonstrate remarkable diversity. Traditional libraries divide their resources into books, periodicals, and so forth; audiovisual materials are located in different sections than are, for instance, manuscripts. The digital library uses a computer screen as its interface, so distinctions of this nature are less important. Many digital libraries can present a variety of materials with equal facility. This overview begins with systems offering multimedia materials, then goes on to discuss specific categories, such as digital books, photographs, and ephemera.

Multimedia Collections

The major multimedia collection, and doubtless the most important digital library, is the American Memory Project, the Historical Collections of the National Digital Library, maintained by the Library of Congress (http://memory.loc.gov/). This now includes over 100 collections, a total of seven million digital items, including photographs, manuscripts, rare books, maps, sound recordings, and moving pictures. The foundations for this were laid as early as 1982–1987, with the production of optical discs. A 1989–1994 program was originally based on CD-ROMs, but it happened to coincide with the beginning of the WWW (Arms, 1996). The year 1994 marked the beginning of the National Digital Library Program (NDLP). Many important collections became available toward the end of that decade. Much digitization was undertaken as collaborative efforts between the Library of Congress and major U.S. universities. Seventeen such collections were awarded development funds; the collections are listed at http://memory.loc.gov/ammem/award/online.html. A typical example is the Traveling Culture collection from the University of Iowa (Figure 3; http://memory.loc.gov/ammem/award98/iauhtml/tccchome.html). This preserves 8,000 brochures

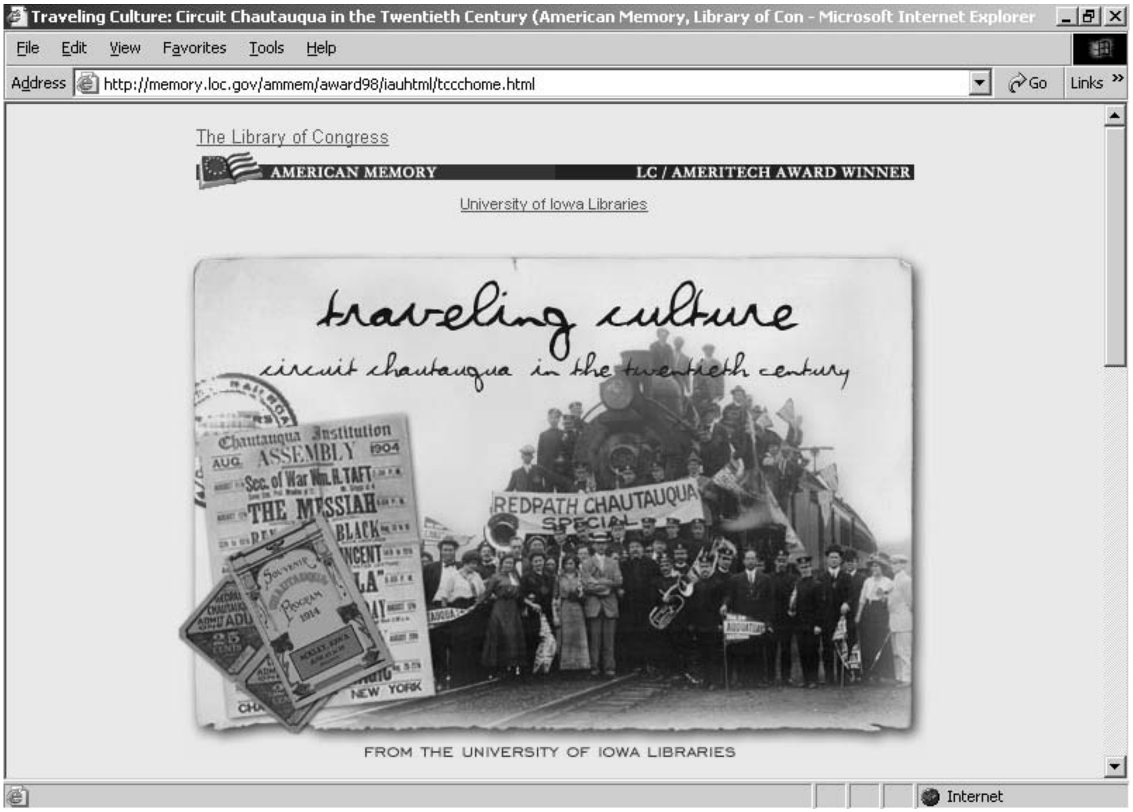

Traveling Culture: Circuit Chautauqua in the Twentieth Century (American Memory, Library of Con - Microsoft Internet Explorer
File
Edit
View
Favorites
Tools
Help
Address
http://memory.loc.gov/ammem/award98/iauhtml/tccchome.html
Go
Links »
The Library of Congress
AMERICAN MEMORY
LC / AMERITECH AWARD WINNER
University of Iowa Libraries
traveling culture
circuit chautauqua in the twentieth century
REDPATH CHAUTAUQUA SPECIAL
FROM THE UNIVERSITY OF IOWA LIBRARIES
Internet

Figure 3: Materials relating to Circuit Chautauqua, lecturers and performers that traveled from city to city in the early 20th century, from a collection at the University of Iowa.

used to publicize Circuit Chautauqua, a system that sent lecturers and serious performers from city to city in the midwest in the early 20th century. Ephemeral materials of this nature are essential for a rounded view of cultural life, but access to the originals was highly restricted. In an arrangement typical of these collaborative efforts, collection details and thumbnail images are loaded by the Library of Congress while the user who requests a full-size image is switched seamlessly to a University of Iowa libraries server.

Several specific collections from the American Memory system are discussed in relevant subject categories below.

Another source of multimedia digital libraries is the Berkeley Digital Library SunSITE (http://sunsite.berkeley.edu/), a collaborative initiative by the Library of the University of California at Berkeley and Sun Microsystems. Their Jack London Collection (http://sunsite.berkeley.edu/London/) offers multiple approaches to this multifaceted author. The collection includes the full text of many of London's books, reproductions of letters and documents, photographs, and audio. It is further supported by a biography, a list of frequently asked questions, listserv, recommended links, and so forth.

The Florida Center for Instructional Technology of the University of South Florida produced the Teacher's Guide to the Holocaust (Figure 4; http://fcit.coedu.usf.edu/Holocaust/), a multimedia collection with a strong message. This offers the text of primary source materials, photographs, films, and music, backed up by maps, bibliographies, and glossaries.

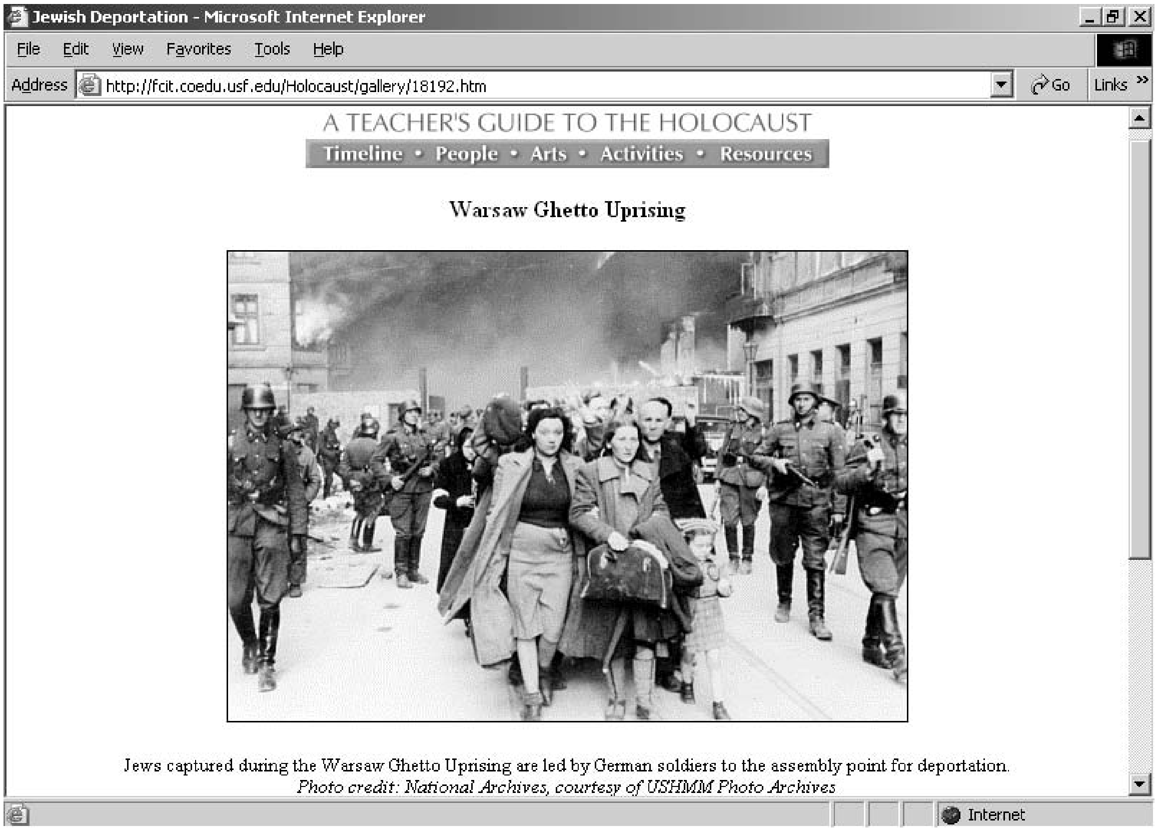
As disk space continues to become more affordable and more bandwidth is available, we can expect to see more multimedia collections.

Book-Oriented Collections

Early digital libraries concentrated on books and were mostly initiated at North American universities. Project Gutenberg (http://promo.net/pg/ or http://gutenberg.net/) is often considered the first significant digital library. It was founded in 1971 by Michael Hart, who was given time on the University of Illinois mainframe computer. The first text distributed was the *Declaration of Independence*, which Michael Hart personally keyboarded. Project Gutenberg has always concentrated on simple presentations, originally FTP downloads of "plain vanilla" ASCII texts, which can be read on almost any computing platform. Many of Project Gutenberg's more than 4,000 texts are now available in HTML, and the volunteer-supported project has numerous mirror sites.

The University of Virginia Electronic Text Center (http://etext.lib.virginia.edu), founded in 1992, now offers 70,000 humanities texts, which have been downloaded more than six million times. Nearly 10,000 of the texts are modern books in English. They are divided into subject groupings, such as the Salem Witch Trials (Figure 5; http://etext.lib.virginia.edu/salem/witchcraft/).

The Center also offers much valuable information on digitization, and links to other English-language full-text Web resources (http://etext.lib.virginia.edu/eng-oth.html). It is one of the few digital libraries with



http://fcit.coedu.usf.edu/Holocaust/gallery/18192.htm

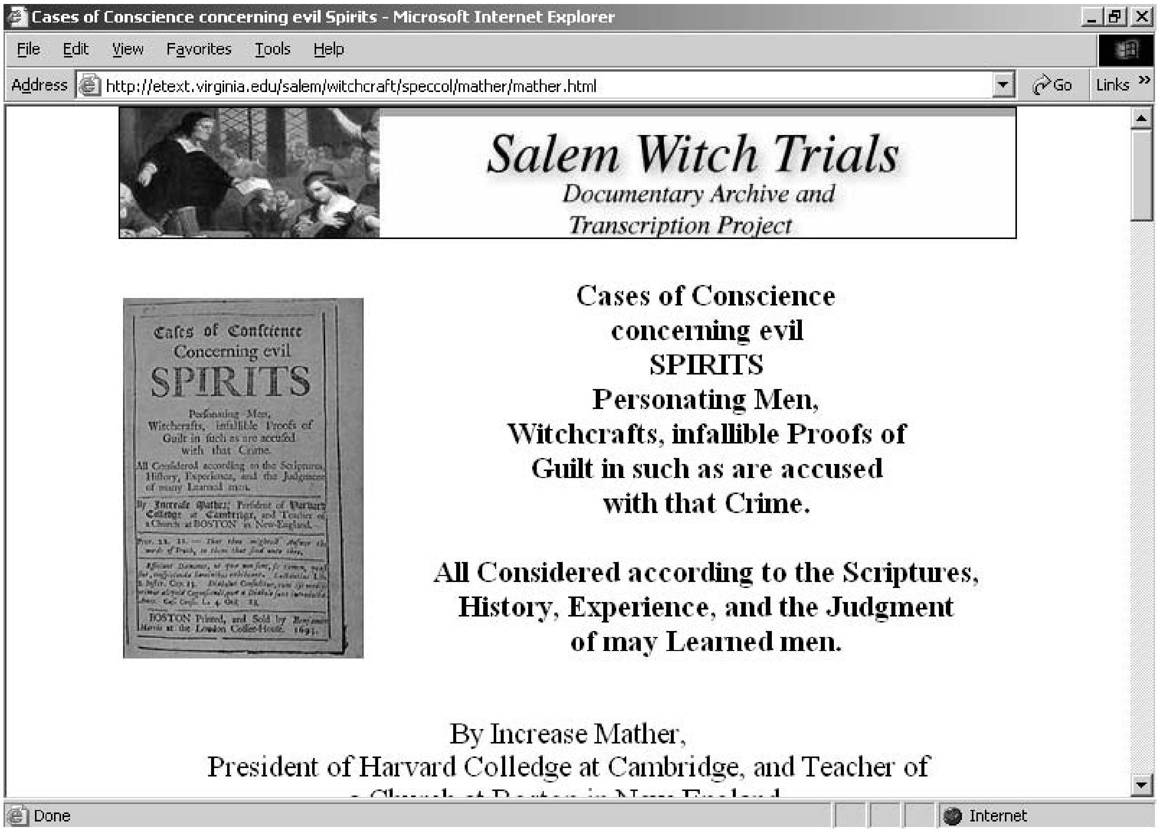
A TEACHER'S GUIDE TO THE HOLOCAUST

Timeline • People • Arts • Activities • Resources

Warsaw Ghetto Uprising

Jews captured during the Warsaw Ghetto Uprising are led by German soldiers to the assembly point for deportation.
Photo credit: National Archives, courtesy of USHMM Photo Archives

Figure 4: Photo taken during the Warsaw Ghetto Uprising from *A Teacher's Guide to the Holocaust*.


Cases of Conscience concerning evil Spirits - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://etext.virginia.edu/salem/witchcraft/speccol/mather/mather.html
Go
Links »
Salem Witch Trials
Documentary Archive and
Transcription Project
Done
Internet

Cases of Conscience
concerning evil
SPIRITS
Personating Men,
Witchcrafts, infallible Proofs of
Guilt in such as are accused
with that Crime.

All Considered according to the Scriptures,
History, Experience, and the Judgment
of may Learned men.

By Increase Mather,
President of Harvard Colledge at Cambridge, and Teacher of

Figure 5: Materials relating to the Salem witch trials, from the University of Virginia Electronic Text Center.

significant offerings in the area of downloadable books, for Palm Pilots or MS Reader (http://etext.lib.virginia.edu/ebooks/ebooklist.html). This is clearly an area with huge potential, as the 1,800-volume collection attracts approximately 300,000 downloads monthly.

The University of Michigan has played a major role in digital library and electronic journal activities. It is responsible for one of the most significant book-oriented projects, the Making of America (http://moa.umdl.umich.edu/). This concentrates on primary source materials for 19th-century American social history from the antebellum period through reconstruction. It contains 8,500 books and 50,000 journal articles, which sophisticated software presents in a choice of image, text, or .pdf files, in normal, large, and small size. The University of Michigan also offers further texts, outside the scope of the Making of America collection, from its Humanities Text Initiative (http://www.hti.umich.edu/index-all.html). Of special interest here is the American Verse Project, preserving pre-1920 American poetry (http://www.hti.umich.edu/a/amverse/).

The American Memory collections of the Library of Congress do include books, but these are mostly specialized collections, such as the 170 Sunday school books published in America between 1815 and 1865, from Michigan State University Libraries and Central Michigan University Libraries (Figure 6; http://memory.loc.gov/ammem/award99/miemhtml/svyhome.html).

A similar specialized collection is the 200 dance instruction manuals digitized at http://memory.loc.gov/ammem/dihtml/dihome.html. The major strength of American Memory is not in the number of books digitized but in the depth and quality of its special collections.

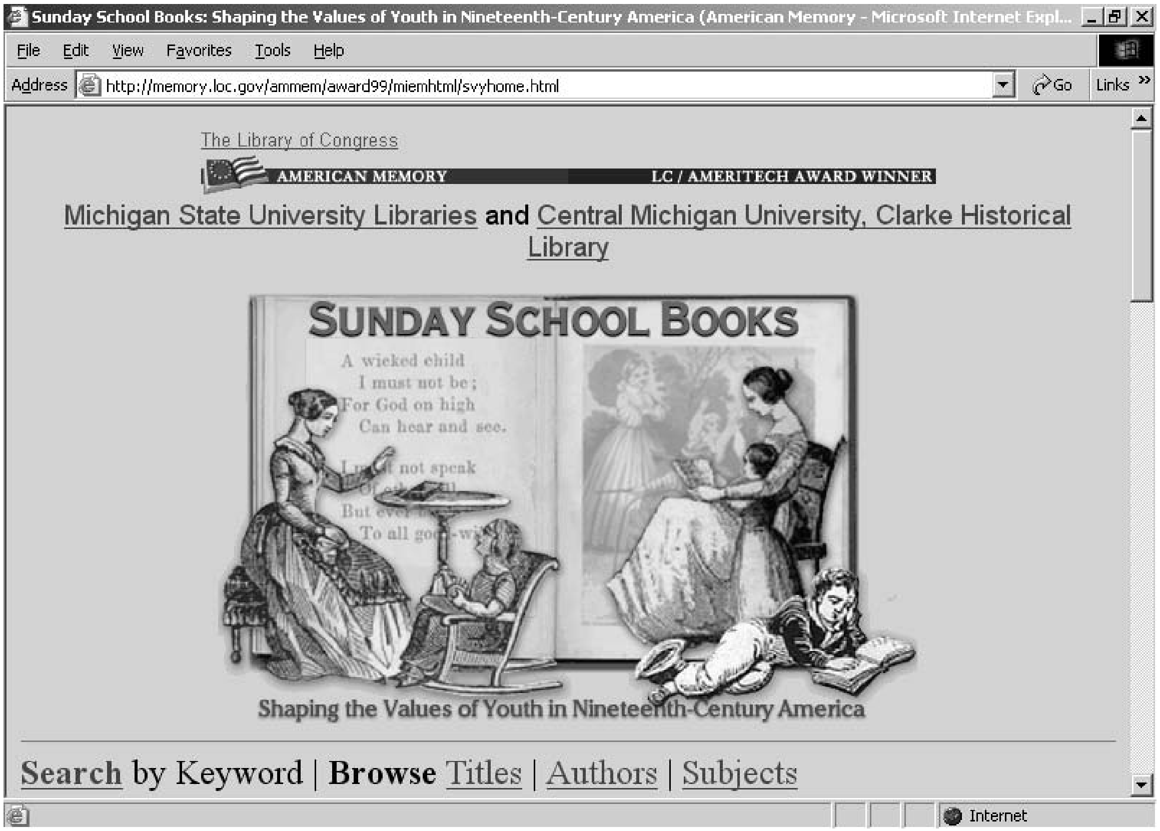
The motto of Bartleby.com is “Great books online” (http://www.bartleby.com/). It offers the classics of English and other literatures and is also strong in authoritative reference materials. It was named after a character from Herman Melville and was founded as a personal research project by an enthusiast in 1993. The first book published was Whitman’s Leaves of Grass.

The Alex Catalogue of Electronic Texts is a collection of public domain documents from American and English literature as well as Western philosophy (http://www.infomotions.com/alex/). It began as a gopher service in 1994.

The English server (http://eserver.org/), founded in 1990 at Carnegie Mellon University, now hosted at the University of Washington, offers over 30,000 texts, including articles, academic resources, and poems. In an unusual collaborative arrangement, most materials are produced or edited by nearly 300 members. Materials are divided into over 40 subject categories and texts may be available in either ASCII or HTML. Resources also include 40 full-length books by major authors.

Image-Oriented Collections

The largest and most important collection of historical photographic images is that devoted to Farm Security Administration photographs (Figure 7; http://memory.



Sunday School Books: Shaping the Values of Youth in Nineteenth-Century America (American Memory - Microsoft Internet Expl...

Address http://memory.loc.gov/ammem/award99/miemhtml/svyhome.html

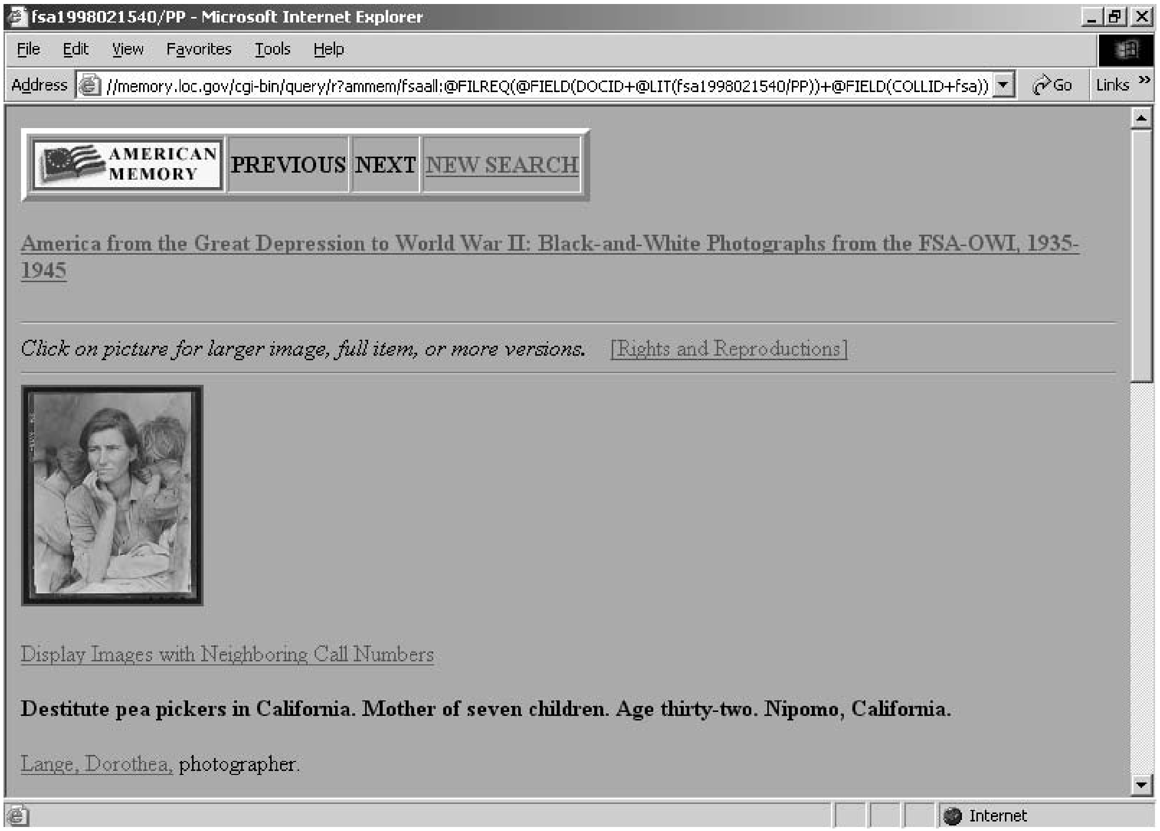
The Library of Congress

AMERICAN MEMORY LC / AMERITECH AWARD WINNER

Michigan State University Libraries and Central Michigan University, Clarke Historical Library

Search by Keyword | Browse Titles | Authors | Subjects

Figure 6: Sunday school books digitized as part of the Library of Congress American Memory Historical Collection.



fsa1998021540/PP - Microsoft Internet Explorer

Address //memory.loc.gov/cgi-bin/query/r?ammem/fsaall:@FILREQ(@FIELD(DOCID+@LIT(fsa1998021540/PP))+@FIELD(COLLID+fsa))

America from the Great Depression to World War II: Black-and-White Photographs from the FSA-OWI, 1935-1945

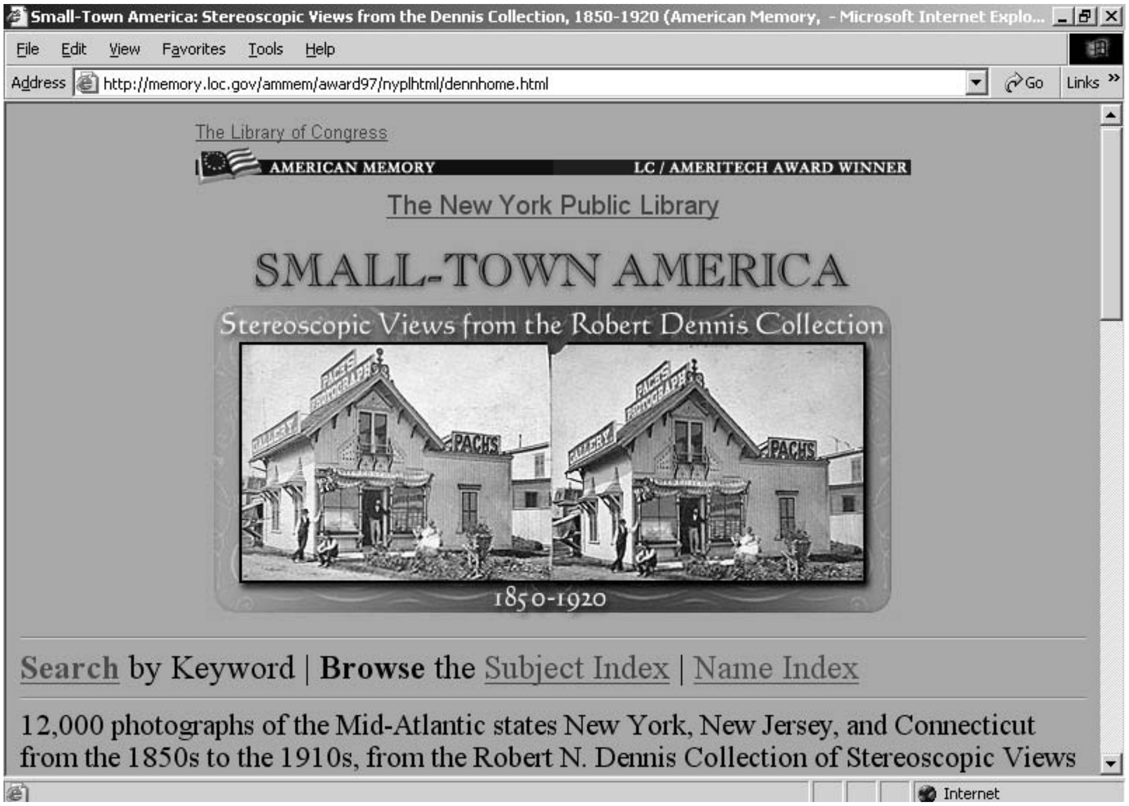
Click on picture for larger image, full item, or more versions. [Rights and Reproductions]

Display Images with Neighboring Call Numbers

Destitute pea pickers in California. Mother of seven children. Age thirty-two. Nipomo, California.

Lange, Dorothea, photographer.

Figure 7: Photograph from the Farm Security Administration collection at the Library of Congress. Taken by Dorothea Lange, this photograph is often referred to as "the Migrant Madonna."



Small-Town America: Stereoscopic Views from the Dennis Collection, 1850-1920 (American Memory, - Microsoft Internet Explo...

http://memory.loc.gov/ammem/award97/nyplhtml/dennhome.html

The Library of Congress

AMERICAN MEMORY LC / AMERITECH AWARD WINNER

The New York Public Library

SMALL-TOWN AMERICA

Stereoscopic Views from the Robert Dennis Collection

1850-1920

Search by Keyword | Browse the Subject Index | Name Index

12,000 photographs of the Mid-Atlantic states New York, New Jersey, and Connecticut from the 1850s to the 1910s, from the Robert N. Dennis Collection of Stereoscopic Views

Figure 8: Stereoscopic views were popular a hundred years ago, and are again available after digitization by the New York Public Library.

loc.gov/ammem/fsowhome.html). This contains about 160,000 black and white photographs, focusing on American rural life during the Depression. Keyword, subject, creator, and geographic indexes are available.

Earlier, equally fascinating, views can be found in "Small-Town America" (Figure 8): stereoscopic views of the mid-Atlantic states, from the 1850s to the 1910s, from the New York Public Library (http://memory.loc.gov/ammem/award97/nyplhtml/dennhome.html).

The stereoscopic views are available in both front and back views. It is especially interesting that the advanced technology of a century ago, stereoscopic views, is being preserved by today's most advanced communication medium, the Internet.

The Bancroft Library at the University of California, Berkeley, offers the California Heritage Collection. This is an online archive of more than 30,000 images illustrating California's history and culture (http://sunsite.berkeley.edu/CalHeritage/).

Manuscripts

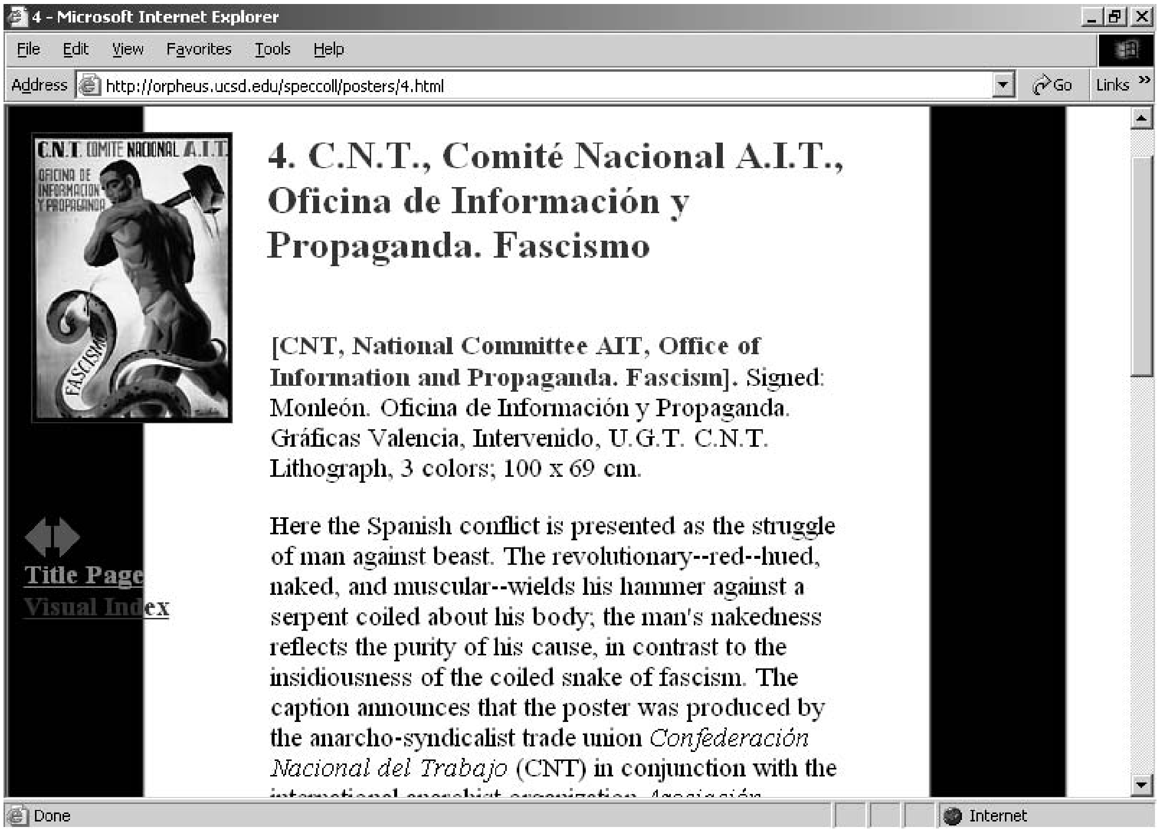
Digital libraries have led to a blurring of the previously strict line between archives and libraries. In fact digital libraries often include significant content of historical manuscripts. The American Memory collection includes the papers of George Washington at the Library of Congress (http://memory.loc.gov/ammem/gwhtml/gwhome.html), a total of 60,000 items. Abraham Lincoln's papers are also available, in approximately 61,000 images and 10,000 transcriptions (http://memory.loc.gov/ammem/alhtml/malhome.html).

Ephemeral Materials

The Internet is an excellent medium for ephemeral materials, so it is no surprise that digital libraries frequently preserve and disseminate ephemera. A well-organized example is the Broadside ballads collection of one of Britain's oldest libraries, the Bodleian Library of Oxford University (http://www.bodley.ox.ac.uk/ballads/). Three thousand items of Historic American Sheet Music from the collections of Duke University can be found at http://memory.loc.gov/ammem/award97/ncdhtml/hasmhome.html. Spanish Civil War posters were digitized at the University of California at San Diego (Figure 9; http://orpheus.ucsd.edu/speccoll/posters/table.html).

Audio and Motion Picture Files

Simple audio and visual files are already available within digital libraries, a trend that, no doubt, will increase rapidly in the future. The section above on multimedia digital libraries gave several examples of systems, which offer partial audiovisual content. A major site dedicated to audio files is History and Politics Out Loud (http://www.hpol.org/). There it is possible to hear a variety of political rhetoric, from Franklin Roosevelt's speech asking Congress to declare war on Japan, to Bill Clinton trying to explain away his relationship with Paula Jones.


4 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://orpheus.ucsd.edu/speccoll/posters/4.html
Go Links
C.N.T. COMITE NACIONAL A.I.T.
OFICINA DE INFORMACION Y PROPAGANDA
FASCISMO
Title Page
Visual Index
Done
Internet

4. C.N.T., Comité Nacional A.I.T., Oficina de Información y Propaganda. Fascismo

[CNT, National Committee AIT, Office of Information and Propaganda. Fascism]. Signed: Monleón. Oficina de Información y Propaganda. Gráficas Valencia, Intervenido, U.G.T. C.N.T. Lithograph, 3 colors; 100 x 69 cm.

Here the Spanish conflict is presented as the struggle of man against beast. The revolutionary--red--hued, naked, and muscular--wields his hammer against a serpent coiled about his body; the man's nakedness reflects the purity of his cause, in contrast to the insidiousness of the coiled snake of fascism. The caption announces that the poster was produced by the anarcho-syndicalist trade union *Confèderación Nacional del Trabajo* (CNT) in conjunction with the international anarchist organization *Asociación*

Figure 9: Spanish Civil War poster, digitized at the University of California, San Diego.

Perhaps the most fascinating cinematographic content in digital libraries is found in the Early Motion Pictures collection of the American Memory collection: (http://memory.loc.gov/ammem/vshtml/vsfilm.html). This preserves and disseminates numerous short films from the dawn of the cinema.

A series of nearly a thousand films is freely available from the Internet Archive (http://www.archive.org/movies/index.html). These are mostly short films, illustrating various aspects of American life in the 20th century.

Electronic Journals (E-journals)

Scientific and professional journals have long been vital elements of library collections. Their electronic equivalents, e-journals, became an area where digital penetration has been intense, and often both commercially driven and controversial. The field can be divided into three major categories: commercial, semicommercial, and free e-journals.

The commercial e-journal field has roots going back to the second half of the 20th century, when the journal field came to be dominated by a small number of specialized publishing houses. They gained a reputation for constantly increasing prices in excess of inflation. As faculty and researchers considered journals essential for research, libraries were under pressure to keep subscribing to ever more expensive periodicals. This had a negative impact on other library areas, notably, book budgets.

Any ideas that the introduction of e-journals would automatically lead to a decrease in journal prices were soon dispelled. Major publishers now offer both paper and electronic versions of their journals; there is often little cost advantage in subscribing to electronic versions alone. In fact, libraries may end up subscribing to bundles of e-journals, including titles in which they have little interest. Prices for the bundles increase, occupying ever larger portions of library budgets. Cancellation of individual titles brings little relief, and librarians complain that they are unable to publicly discuss their contract terms (Foster, 2002). Access to electronic versions of established journals is limited to identifiable members of universities and similar institutions and is controlled by a small number of established publishers and subscription agents. These closed-access systems have little impact on the Internet in general. Major systems include Elsevier, a leading name in scientific journal publishing, which maintains Science Direct (http://www.sciencedirect.com/), offering controlled electronic access to 1,700 titles, and 3.9 million articles. Blackwell's Synergy system permits nonsubscribers to purchase single items (http://www.blackwell-synergy.com/). Ebsco, long-established as an agent fulfilling library journal subscriptions, now permits libraries to access over eight 8,000 e-journals (http://www.ebsco.com/ess/services/online.stm). Because of the volatility of e-journals and the potential loss of long-term electronic access, libraries prefer to purchase e-journal subscriptions via established sources.

This situation has come to be called the Serials Crisis by faculty and librarians, whose opinions are disseminated in Create Change (http://www.arl.org/create/home.html). This site states that the 121 members of the Association of Research Libraries spend about $480 million per year on journal collections, an average of about $4 million each. A consortium of 80 Ohio academic libraries was spending about $19 million annually for access to about 4,500 journals (Foster, 2002). Faculty and researchers write journal articles and donate them to journal publishers, normally without receiving royalties. They also generally work free of charge as referees and editors for scientific journals. These same journals are then sold back to the libraries of their institutions at a steep price. Electronic journals are seen as a way out of this situation. One solution is offered by SPARC, the Scholarly Publishing and Academic Resources Coalition (http://www.arl.org/sparc/home/index.asp?page = 0), which has set up a collaborative system for academic e-journal production. Intermediate, semicommercial approaches are offered by Project Muse, of Johns Hopkins University Press (Figure 10; http://www.press.jhu.edu/muse.html), and JSTOR, the scholarly journal archive (http://www.jstor.org/), organized by the University of Michigan. Project Muse offers reasonably priced access to over 100 titles. JSTOR works on a nonprofit basis and handles backfiles of over 200 titles. They both limit access to members of subscribing institutions.

The International Consortium for the Advancement of Academic Publication, based at Athabasca University, Canada (http://www.icaap.org/), offers a collaborative system for publishing specialized e-journals.

The field of free e-journals has also expanded rapidly. The Internet Library of Early Journals, a cooperative UK program, offers free access to six key 18th and 19th century journals (http://www.bodley.ox.ac.uk/ilej/). A significant portion of the e-journal field can be searched database style via the Electronic Journal Miner set up by the Colorado Alliance of Research Libraries (http://ejournal.coalliance.org/). Their site also lists other guides to e-journals (http://ejournal.coalliance.org/info/other.html). Three thousand titles are listed at the Internet Public Library (http://www.ipl.org/reading/serials/). An additional list of e-journals can be found at the World Wide Web Virtual Library (http://www.e-journals.org/). The Association of Research Libraries published the first printed list of e-journals in 1991, with 173 pages. The latest edition of its Directory of Scholarly Electronic Journals and Academic Discussion Lists (2000) has over 1,100 pages and includes nearly 4,000 peer reviewed journals (http://www.arl.org/scomm/edir/index.html).

Theoretically, electronic journals offer additional advantages in relation to traditional journals because they can incorporate audio, music, short films, animations, and so forth. So far, this potential has been underexplored. Animated displays of physical data can be found in a SPARC journal, the *New Journal of Physics* (go to http://www.njp.org and look for papers which include multimedia elements).

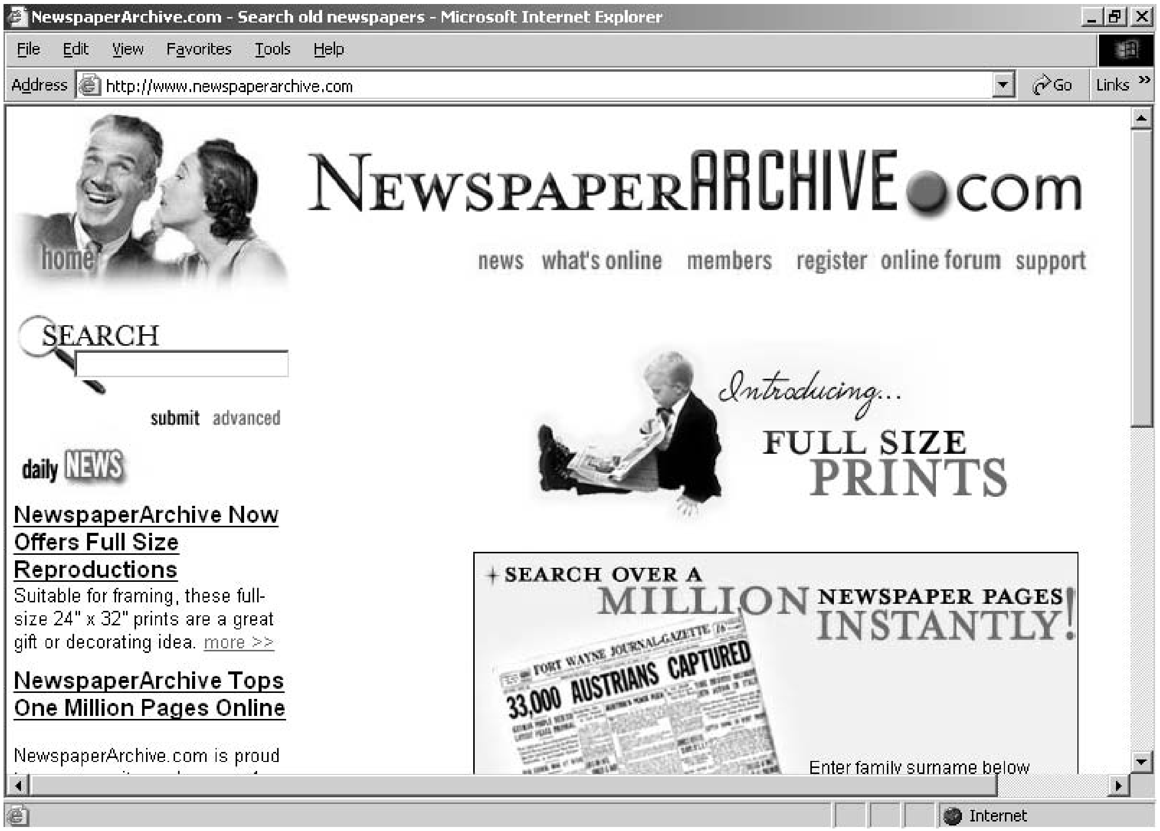
Newspapers present difficult problems of storage, analysis, and retrieval of information. NewspaperArchive (Figure 11) is a well-indexed source: Five newspapers are

Project MUSE - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://muse.jhu.edu/
Go
Links
Project MUSE®
Scholarly Journals Online
Journals
Subscribe
Search
Information
What's New
Contact
Publishers
Internet

Project MUSE provides online, worldwide, institutional subscription access to the full text of more than 200 scholarly journals in arts and humanities, social sciences and mathematics.

THE JOHNS HOPKINS UNIVERSITY PRESS in collaboration with THE MILTON S. EISENHOWER LIBRARY

Figure 10: Project MUSE, from Johns Hopkins University Press, permits access to over 200 digitized scholarly journals.



NewspaperArchive.com - Search old newspapers - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address
Go
Links
NEWSPAPERARCHIVE.com
home
news what's online members register online forum support
SEARCH
submit advanced
daily NEWS
Introducing...
FULL SIZE PRINTS
SEARCH OVER A MILLION NEWSPAPER PAGES INSTANTLY!
FORT WAYNE JOURNAL-GAZETTE
33,000 AUSTRIANS CAPTURED
Internet

http://www.newspaperarchive.com

NewspaperArchive Now Offers Full Size Reproductions

Suitable for framing, these full-size 24" x 32" prints are a great gift or decorating idea. more >>

NewspaperArchive Tops One Million Pages Online

NewspaperArchive.com is proud

Enter family surname below

Figure 11: NewspaperArchive is a commercial service offering access to historic newspapers.

free; a subscription is charged for full access. In April 2002, this Cedar Rapids company claimed to have over one million pages available online, and to be adding 11,000 to 14,000 pages a day (http://www. newspaperarchive.com/default.asp).

Electronic Theses and Dissertations (ETDs)

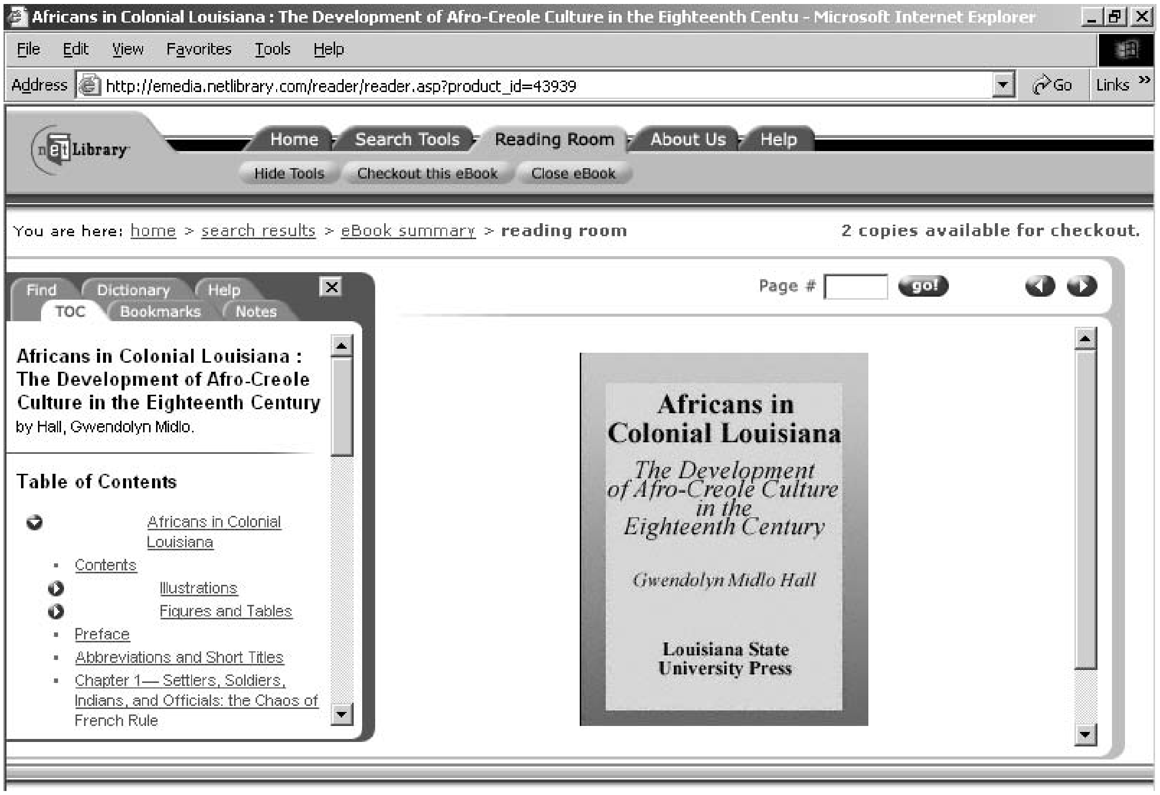
Theses and dissertations are produced in high-tech institutions and universities and are of interest to highly literate users, who are accustomed to retrieving information via the Internet. These are generally produced according to standardized print formats, and they have also traditionally been difficult to obtain. The potential for electronic theses and dissertations is therefore great; constraints seem to be the lack of standardized delivery systems, and the fear of some authors that wider thesis availability might prejudice later book publication. A general reference point is the Networked Digital Library of Theses and Dissertations (NDLTD; http://www.ndltd.org/). A founding member of this was Virginia Tech, which now offers unrestricted access to over 2,000 digital theses (http://scholar.lib.vt.edu/theses/). Experience shows that digital availability stimulates use and promotes the author's reputation. West Virginia University reports over 1.4 million hits on its digital theses collection (http://www.wvu.edu/~thesis/News/WVU_ETDS_Over_1_Million_Served.htm). Usage of digital theses far outpaces consultation of paper equivalents. The "champion" thesis at WVU had been accessed 13,000 times. An interesting intermediate solution is offered by MIT (http://thesis.mit.edu/), which permits reading of theses in a presentation that inhibits online printing; interested persons can buy .pdf versions online, or purchase paper copies which are then sent by mail.

Pointer or Signpost Sites

The first attempt to produce a guide to the WWW was by the creator of the WWW, Tim Berners-Lee, shortly after he created it. The WWW Virtual Library, a cooperatively updated system, can still be found at http://vlib.org/Home.html.

Sites that simply direct users to Internet resources are considered digital libraries when they are organized in a manner similar to libraries. The most significant of these services is the Internet Public Library (http://www.ipl.org), originally set up in 1995 as a class project at the School of Information at the University of Michigan (Magpantay, J. A. et al., 1997). It has sections for reference and texts, magazines and newspapers, teens and youths, and so forth.

Those who really love libraries will enjoy lists of Web resources classified by the Dewey Decimal Classification system (http://www.oclc.org/oclc/fp/worldwide/web_resources.htm or http://bubl.ac.uk/link/ddc.html). BUBL, originally the Bulletin Board for Libraries, also offers subject area access to the Internet (http://bubl.ac.uk/link/). A major British contribution to Internet information retrieval is the creation of subject gateways



Africans in Colonial Louisiana : The Development of Afro-Creole Culture in the Eighteenth Centu - Microsoft Internet Explorer

http://emedia.netlibrary.com/reader/reader.asp?product_id=43939

You are here: home > search results > eBook summary > reading room

2 copies available for checkout.

Africans in Colonial Louisiana : The Development of Afro-Creole Culture in the Eighteenth Century
by Hall, Gwendolyn Midlo.

Table of Contents

- Africans in Colonial Louisiana
- Contents
- Illustrations
- Figures and Tables
- Preface
- Abbreviations and Short Titles
- Chapter 1— Settlers, Soldiers, Indians, and Officials: the Chaos of French Rule

Africans in Colonial Louisiana
The Development of Afro-Creole Culture in the Eighteenth Century
Gwendolyn Midlo Hall
Louisiana State University Press

Figure 12: NetLibrary permits access to digital copies of books of scholarly interest.

for specific subject areas. Probably the most successful of these is the Edinburgh Engineering Virtual Library (http://www.eevl.ac.uk); there is also ADAM (Art, Design, Architecture and Media Information Gateway; http://www.adam.ac.uk), Organizing Medical Networked Information (OMNI; http://omni.ac.uk), and the Social Science Information Gateway (http://sosig.ac.uk). A general resource, linked to these British sites, is the Resource Discovery Network, a gateway to, or a directory of, gateways (http://www. rdn.ac.uk/).

Infomine was founded by librarians at the University of California, Riverside, in 1994 (http://infomine.ucr.edu/). It offers both directory and search access to 23,000 Internet resources. The California State Library maintains the Librarian's Index to the Internet, which includes 6,400 Internet resources (http://lii.org/).

Project i-DLR was developed by faculty and students at the University of Missouri-Columbia (http://whistletest.coe.missouri/~rafee/iDLR/index.php) to provide access to educational materials in support of digital library concepts and practices.

Commercially Oriented Digital Libraries

The history of the Internet, so far, can be divided into two major phases. The first was exploration and initial content creation, led by academics and enthusiasts. This was followed by the arrival of commercial interests, who looked around and asked how they could make money out of this. It is notoriously difficult to make money from library services and so far there has been considerable difficulty in establishing successful, commercially profitable, digital libraries covering a wide range of materials. The situation was worsened by the dot-com shakeout of late 2000 to mid-2001. Probably the most successful system to date is NetLibrary (Figure 12; http://www.netlibrary.com/), which offers 40,000 titles and has always concentrated on selling through libraries. NetLibrary permits simultaneous access to only a limited number of copies of each book. This appears strange in view of the Internet's capability of generating an unlimited number of copies of a digital text, but this restriction encourages publishers to join the system. Rapid printing of multiple pages is also inhibited.

In 2002, NetLibrary was sold to OCLC (Online Computer Library Center; http://www.oclc.org/home/), a cooperative system that supplies automated cataloging data to more than 40,000 libraries. The OCLC's stability and the ease of access to libraries should guarantee a bright future for NetLibrary.

Another major commercial digital library initiative is Houston-based Questia (http://www.questia.com/), which claims to be the world's largest online library of books. It digitized a core collection in the humanities and social sciences, excluding business and science texts, and offers 45,000 complete books and 25,000 journal articles. There is an excellent advanced software interface and full access costs $19.95 a month. The third player in this field is ebrary (http://www.ebrary.com/), financed by major publishers, including Random House and McGraw–Hill. This

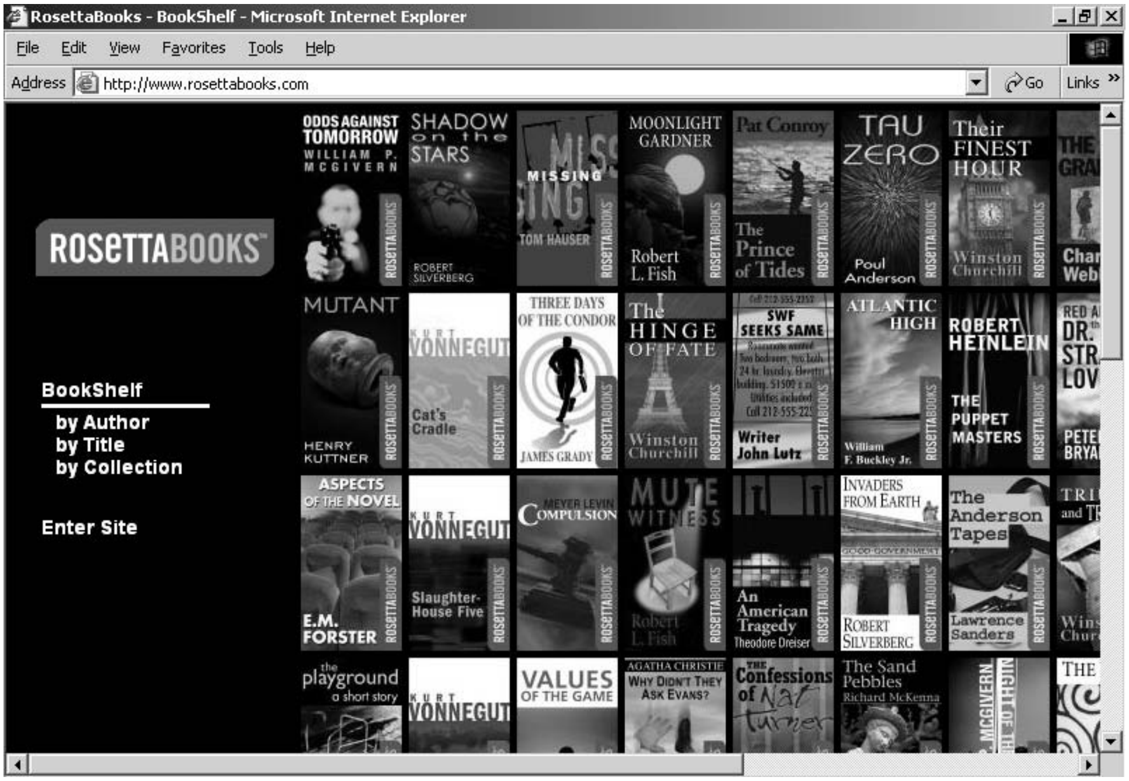

RosettaBooks - BookShelf - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://www.rosettabooks.com
Go
Links
ROSETTABOOKS
BookShelf
by Author
by Title
by Collection
Enter Site
ODDS AGAINST TOMORROW WILLIAM P. MCGIVERN
SHADOW on the STARS ROBERT SILVERBERG
MISSING TOM HAUSER
MOONLIGHT GARDNER Robert L. Fish
Pat Conroy The Prince of Tides
TAU ZERO Poul Anderson
Their FINEST HOUR Winston Churchill
MUTANT HENRY KUTTNER
KURT VONNEGUT Cat's Cradle
THREE DAYS OF THE CONDOR JAMES GRADY
The HINGE OF FATE Winston Churchill
SWF SEEKS SAME Writer John Lutz
ATLANTIC HIGH William F. Buckley Jr.
ROBERT HEINLEIN THE PUPPET MASTERS
ASPECTS OF THE NOVEL E.M. FORSTER
KURT VONNEGUT Slaughter-House Five
MEYER LEVIN COMPULSION
MUTE WITNESS Robert L. Fish
An American Tragedy Theodore Dreiser
INVADERS FROM EARTH ROBERT SILVERBERG
The Anderson Tapes Lawrence Sanders
the playground a short story
KURT VONNEGUT
VALUES OF THE GAME
AGATHA CHRISTIE WHY DIDN'T THEY ASK EVANS?
The Confessions of Nat Turner
The Sand Pebbles Richard McKenna
MCGIVERN

Figure 13: Rosetta Books specializes in the publication of downloadable e-books.

was being launched in 2002 with a business model that permitted consumers to view pages free but charged them to print.

The Association for Computing Machinery (ACM) is an outstanding example of a professional association that has been able to successfully implement a paid-access digital library, the ACM Digital Library (http://portal.acm.org/). It has, of course, a major advantage, in that it is aimed at an audience with a very high level of computer literacy. Niche services offered by professional associations may in some cases offer greater potential than purely commercial systems.

This chapter deals above all with libraries viewed via the Internet, but an important parallel area is springing up: that of commercially produced e-books downloaded to user's devices. The possibilities exploded into public view in the year 2000, when half a million people downloaded Stephen King's e-book, Riding the Bullet. Notable among specialized publishers of downloadable e-books is Rosetta Books (Figure 13; http://www.rosettabooks.com).

Rosetta Books received considerable publicity in 2001 when Random House tried to stop them from distributing electronic versions of works by Kurt Vonnegut, William Styron, and Robert B. Parker. The judge ruled in favor of Rosetta Books, determining that ebooks are a separate medium from the original product, because they offer full-text searching, hyperlinking, and other electronic advantages (Random House, 2001). Palm Digital Media at Peanut Press (http://www.peanutpress.com/) offers 5,000 titles for sale. PalmBooks links to over 1,600 free texts readable on Palm hand-helds (http://www.palmbooks.org/default.asp). For a general overview of downloadable e-books, look at the relevant subdivision of the major online bookshop sites, Barnes and Noble (http://www.bn.com/) or Amazon (http://www.amazon.com/).

PROCEDURES AND PRACTICES IN THE DIGITAL LIBRARY FIELD

Selection and Collection Development

The first step in digital library creation is to identify and select appropriate originals for dissemination in digital form. The basic principles are similar to those of traditional library collection development policy: to select quality materials of high potential demand for the intended audience of end users. In many cases, rare books, photographs, and ephemera from special collections maintained by university libraries will form the nucleus of successful digital libraries (Love, 1998). If a formal project is to be submitted to a funding body, they will need to know as exactly as possible how many items will be scanned and processed. Digital selection, unlike traditional selection, is not constrained by lack of physical space, or by the range of materials available from traditional publishers.

Copyright

Beyond any doubt, the major constraint on the digital library field comes from the copyright laws. These were written to control physical objects and adapt poorly to digital environments. Constitutionally, copyright is granted

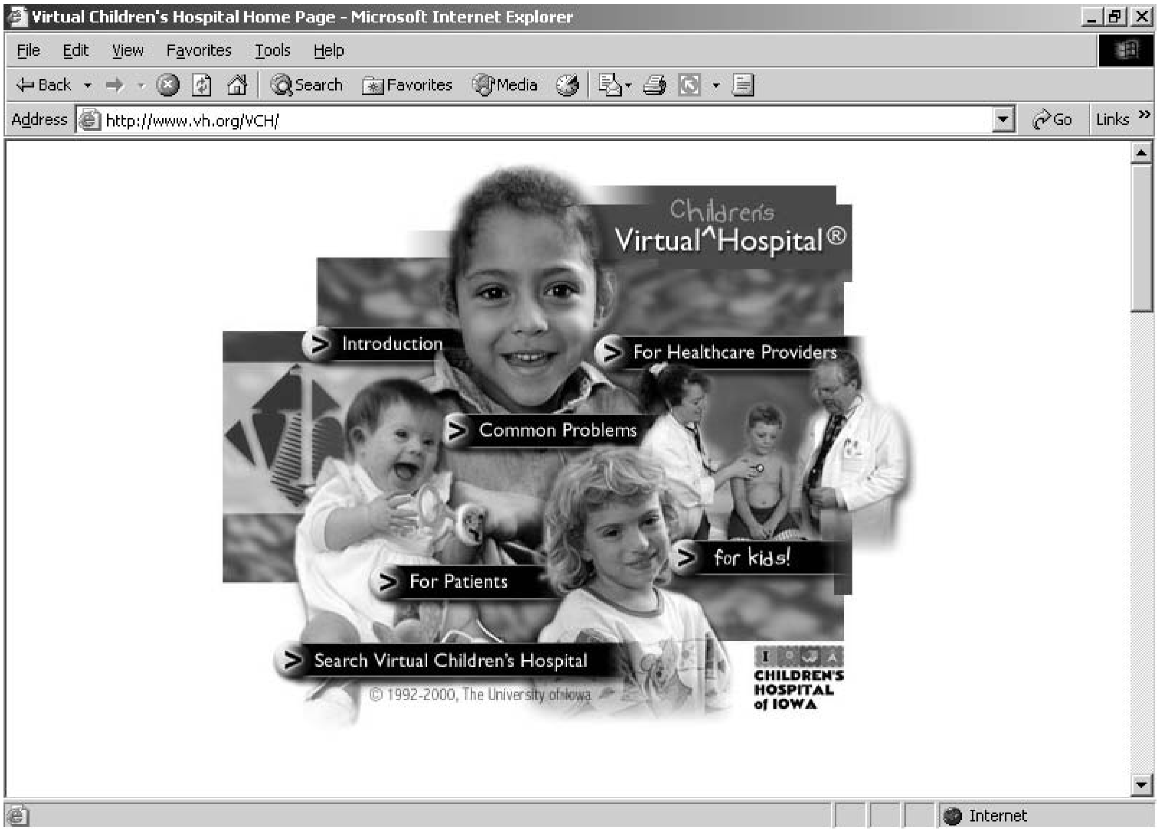

Virtual Children's Hospital Home Page - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.vh.org/VCH/
Go Links
Children's Virtual^Hospital®
Introduction
For Healthcare Providers
Common Problems
for kids!
For Patients
Search Virtual Children's Hospital
© 1992-2000, The University of Iowa
CHILDREN'S HOSPITAL of IOWA
Internet

Figure 14: The Virtual Children's Hospital, maintained by the University of Iowa, offers access to specialized health care information.

for "limited times," but terms have been extended, largely in response to pressures from film studios. The Sonny Bono Copyright Act, 1998, fixed copyright at the life of the author plus 70 years. Pre-1978 works in copyright are protected for 95 years from the date copyright was first secured. Most materials produced in 1923 or after are protected by copyright. Consult the Public Domain and Copyright How-To page of Project Gutenberg (http://promo.net/pg/vol/pd.html) or "When works pass into the public domain" (http://www.unc.edu/~unclng/public-d.htm). Note that all works are protected—books, films, photographs, ephemera, newspapers—anything fixed in tangible form, whether or not anybody is interested in protecting their rights in these materials. In 1930 about 10,000 books were published in the United States; today fewer than 200 are still available from publishers (Brief, 2002). The other 9,800 may well include texts, which might usefully be made available to the public in digital form, but they continue to enjoy full copyright protection until 2019. No institution could risk digitizing them, for fear that rights-holders might suddenly appear.

In many cases, digital libraries find it simpler to limit themselves to pre-1923 materials, to avoid the troublesome process of identifying rights-holders. Most book publishing contracts signed in the United States in the 20th century state that rights return to the author a few years after the book goes out of print (Lynch, 2001). When the authors die, these rights pass to their descendents. This has created a huge class of "orphan books," items for which no rights-holder can be traced. The situation is even less clear for photographs and ephemera. If the rights-holder can be identified, the digital library has to go through a lengthy, often expensive, and sometimes fruitless process of negotiating rights.

One solution is to work with copyright-free niche collections, which avoid copyright problems. The Government Publications from World War II at Southern Methodist University (http://worldwar2.smu.edu) can be freely disseminated, because U.S. government publications are not copyrighted. A similar example comes from the Truman Presidential Library and Museum (http://www.trumanlibrary.org/), which, like many of the Presidential Libraries under direction of the National Archives and Records Administration (NARA), provides access to records that are in the public domain. The University of Iowa maintains medical documentation on its Virtual Hospital and Virtual Children's Hospital sites (Figure 14; http://www.vh.org or http://www.vch.org). These are based on noncopyright government publications and texts contributed by faculty of the University of Iowa.

Also in the medical field, PubMed, set up by the National Library of Medicine, facilitates public awareness of 11 million items of medical literature (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi). There are no copyright problems because they offer abstracts rather than full text.

Another possibility is to concentrate on tribute sites, for which permission is normally readily given. The Jack London tribute site at Berkeley SunSITE has already been cited; SunSITE maintains a similar site for the anarchist and feminist Emma Goldman (http://sunsite.berkeley.edu/Goldman/). This includes selected texts, documents, and even newsreel footage.

Further possibilities are opened up by licensing systems, such as Open Content (http://opencontent.org/index.shtml), which does for publications what Open Source licensing does for software. Open Content is freely available for modification, use, and redistribution; but if someone else bases a work on Open Content, the result must also be made available as Open Content. The license relieves authors of liability or implication of warranty and insures that they will be properly credited when Open Content is used. Others may modify and redistribute the content if they clearly indicate the changes that have been made.

Digitization of Materials for Digital Libraries

High-quality scanning is typically used for images incorporated into digital libraries. It is standard practice to produce three files. First there is an archival or master file normally in .tiff format (tagged image file format); this will be carefully archived but may not be available to the public. Next come the viewing or access files, normally in .jpg format (Joint Photographic Experts Group), although Adobe's .pdf (portable document file) system can also be used. These are currently the most popular formats for general image display. Finally, it is common to produce a thumbnail image, usually a small .jpg file or .gif (graphic interchange format), for quick-reference and table-of-content purposes. A useful summary of current practice can be found in the Arizona State Library, Archives and Public Records' Digitization Guidelines (http://www.lib.az.us/digital/dg_a4.html). These suggest, for photographic reproductions, a resolution of 600 dpi (dots per inch) for archive images, 300 dpi for viewing images, and 72 dpi for thumbnails.

Digital libraries also need to convert page images to text that can be manipulated. This would now normally be HTML text, although early systems used plain ASCII .txt files. Conversion to text is undertaken using quality OCR (optical character recognition) software, such as Omnipage Pro (University of Virginia Electronic Text Center, 1998). Even first-class scanning software can confuse certain words, for example clean/dean or modern/modem. Careful proofreading is therefore essential. The problem is especially acute in digital libraries, which pride themselves on offering quality content but which process a large proportion of older and specialized texts, with rarely used terms, unusual spellings, unusual fonts, and so forth. Highly qualified personnel are required to proofread such texts.

Text Indexing

Complete, word-by-word text indexing is a valuable feature of most digital libraries. Specialized software for this area includes the XPAT search engine, part of the University of Michigan Digital Library eXtension Service (DLXS) software suite (http://www.dlxs.org/). This software can be seen in action in one of the most daunting indexing challenges faced by a digital library, the Legacy National Tobacco Documents Library (LNTDL; http://legacy.library.ucsf.edu/index.html). The collection contains over 20 million documents, obtained through the legal discovery process for a lawsuit, relating to scientific research, manufacturing, marketing, advertising, and sales of cigarettes.

Another specialized example is the Simple Web Indexing System for Humans-Enhanced (SWISH-E; http://swish-e.org/). This Open Source software is used to enable full-text indexing of many sites, including the Berkeley Digital Library SunSITE (http://sunsite.berkeley.edu/cgi-bin/search.pl), and even the Apache Web Server site (http://search.apache.org/).

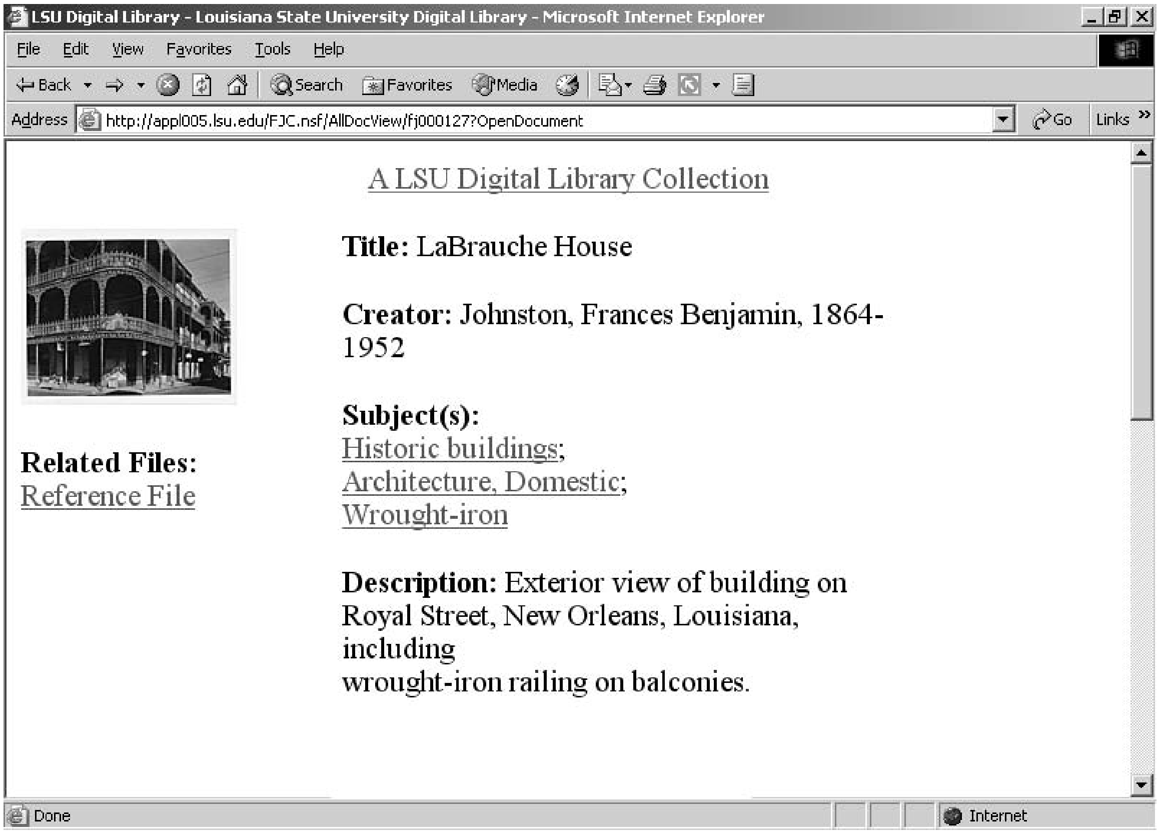
Metadata

The addition of Metadata or subject terms adds significant value to digital records. The American Memory collections are especially notable for enriching records with numerous Library of Congress subject headings. The best-known system, specifically developed for a digital environment, is the Dublin Core, first developed at a meeting in Dublin, Ohio, in 1995. The Dublin Core Metadata Element Set (http://dublincore.org/documents/dces/) specifies metadata elements such as title, creator, subject, and rights (for digital rights management information).

The use of metadata is clearly demonstrated in Figure 15, from the Louisiana State University Digital Library (http://www.lsu.edu/diglib/).

The source code for the page, containing the photograph of a famous building in the French Quarter of New Orleans (http://APPL005.lsu.edu/FJC.nsf/AllDocView/fj000127?OpenDocument), includes the following metadata; "DC" identifies this as Dublin Core metadata.

```
<META name="DC.Title" content="LaBrauche
  House"
<META name="DC.Creator" content="Johnston,
  Frances Benjamin, 1864-1952"
<META name="DC.Subjects" content="Historic
  buildings";"Architecture, Domestic";
  "Wrought-iron";
<META name="DC.Description"
  content="Exterior view of building on
  Royal Street, New Orleans, Louisiana,
  including wrought-iron railing on
  balconies."
<META name="DC.Type" content="image"
<META name="DC.Format" content="jpeg"
<META name="DC.Identifier" content="http://
  APPL005.lsu.edu/FJC.nsf/AllDocView/
  fj000127?OpenDocument"
<META name="DC.Source" content="Louisiana
  State Museum"
<META name="DC.Language" content="en"
<META name="DC.Coverage" content="French
  Quarter (New Orleans, La.)"
<META name="DC.Rights" content="Physical
  rights are retained by the Louisiana
  State Museum. Copyright is retained in
  accordance with U.S. copyright laws."
```


LSU Digital Library - Louisiana State University Digital Library - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address
Go Links
Done
Internet

http://appl005.lsu.edu/FJC.nsf/AllDocView/fj000127?OpenDocument

A LSU Digital Library Collection

Title: LaBrauche House

Creator: Johnston, Frances Benjamin, 1864-1952

Subject(s):
Historic buildings;
Architecture, Domestic;
Wrought-iron

Description: Exterior view of building on Royal Street, New Orleans, Louisiana, including
wrought-iron railing on balconies.

Related Files:
Reference File

Figure 15: Louisiana State University has digitized historic photographs from the Louisiana State Museum.

Metadata creation requires highly trained analysts, similar to traditional catalogers, normally working with a list of acceptable terms. It can therefore add significantly to both the value and the cost of systems.

XML: Extensible Markup Language

Digital libraries often employ a markup language (such as extensible markup language, XML), to give additional structural meaning to the text. The best-known markup language, hypertext markup language (HTML), the basis of the WWW, defines the appearance of specific blocks of text, which can appear as bold, centered, and so forth. XML uses a similar system of tags in pointed brackets, but it can specify semantic elements. It can identify a specific piece of text as being an author's name, an article title, an abstract, and so forth. XML is better thought of as a toolkit for developing a markup language, whereas HTML is a simple markup language with fixed capabilities. HTML basically controls presentation and is not extensible. One of the most easily consulted examples of XML in a digital library context is that created by the National Library of Medicine to facilitate communication of abstracts of medical literature to PubMed. An extract is presented below.

```
<Journal>
<PublisherName>AAAS</PublisherName>
<JournalTitle>Science</JournalTitle>
<Volume>271</Volume>
<Issue>5</Issue>
<PubDate>
<Year>1996</Year>
<Month>Mar</Month>
<Day>3</Day>
</PubDate>
</Journal>
```

This abbreviated example is for the journal *Science*, volume 271, issue 5. The full example, which includes authors and their addresses, title, and abstract of the paper, can be consulted at http://www.ncbi.nlm.nih.gov:80/entrez/query/static/spec.html#ExampleXMLfile.

The use of XML and other markup languages in digital libraries is a complex subject; a useful introductory overview of the field can be found on the University of Virginia's Electronic Text Center's Standards page (http://etext.lib.virginia.edu/standard.html). This also covers other systems used in the field, such as Standard General Markup Language (SGML), Text Encoding Initiative (TEI), and Electronic Archival Description (EAD).

Computing Systems

Special-purpose software solutions include the University of Michigan DLXS software suite (http://www.dlxs.org/), which can be seen in action in Michigan's Making of America collection (http://moa.umdl.umich.edu/).

The Greenstone Digital Library software is an interesting free Open Source software suite for digital libraries (http://www.greenstone.org/english/home.html). It

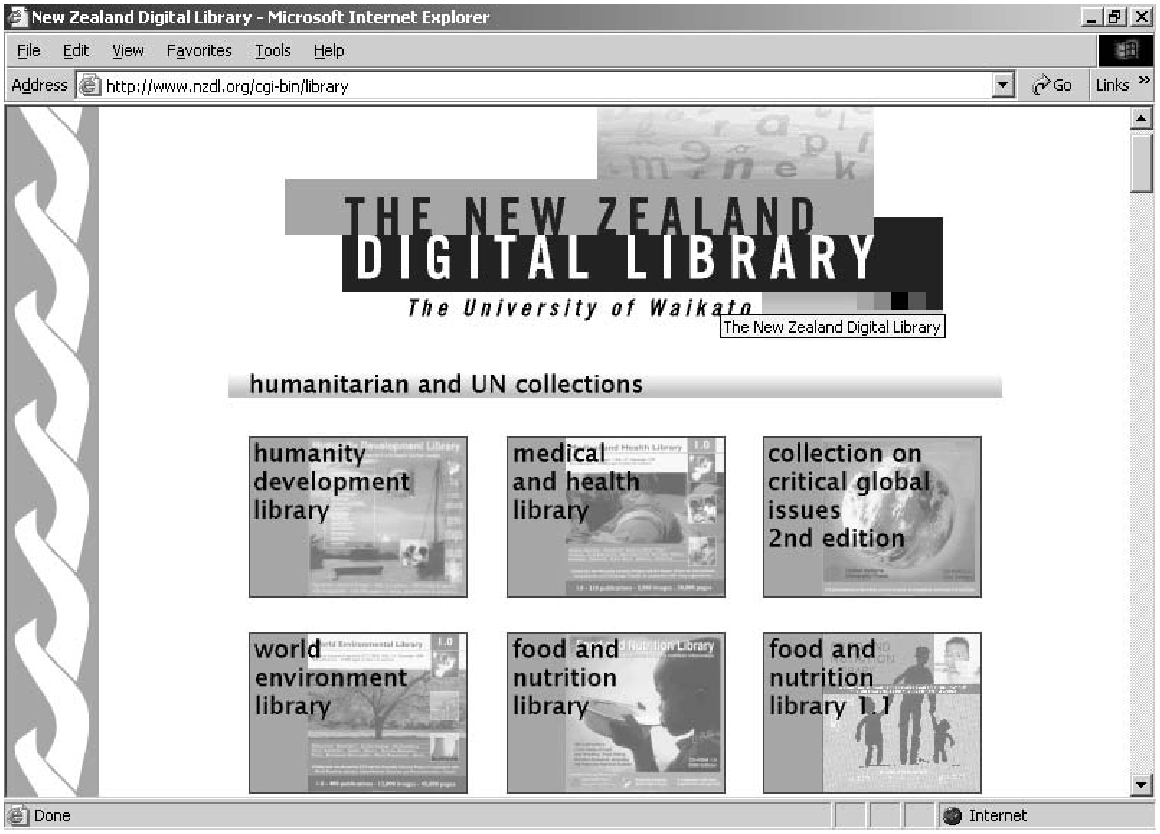

New Zealand Digital Library - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://www.nzdl.org/cgi-bin/library
Go
Links
THE NEW ZEALAND
DIGITAL LIBRARY
The University of Waikato
The New Zealand Digital Library
humanitarian and UN collections
humanity development library
medical and health library
collection on critical global issues 2nd edition
world environment library
food and nutrition library
food and nutrition library 1.1
Done
Internet

Figure 16: The New Zealand Digital Library offers access to documents relevant to developing countries.

was developed at a university in New Zealand, and its major application is the New Zealand Digital Library (Figure 16; http://www.nzdl.org/cgi-bin/library). It offers multiscript capabilities and has also been used for collections of United Nations documents relevant to developing countries.

Open Source software is widely used for server software and in academic environments. It is therefore natural that it is often preferred for digital library database work. The best known standardized language for databases is SQL (Structured Query Language).

In the digital library field, it is normally applied via Open Source server-side database software, of which PostgreSQL and MySQL are frequently cited as powerful alternatives. A combination of MySQL with PHP has drawn much attention recently. PHP (http://www.php.net) is a server-side Web-scripting language whose predefined function set easily interfaces with MySQL. The Digital Library Toolkit (Noerr, 2000) contains a useful overview of software. ColdFusion is a commercial database software alternative.

Z39.50 compliant software will be required to query traditional library catalogs and bibliographic databases (*Biblio Tech Review*, 2001). It is often considered too complex for large-scale application in an Internet environment. One possible solution is the Open Archives Initiative (OAI; http://www.openarchives.org/), which is based on Harvester/Server Architecture. Participating data stores would make simple metadata available in standardized format. Information systems would harvest relevant metadata, which would then be used to power relevant subject gateways. The model is being tested with Americana, environmental information, and academic publications (Digital Library Federation, 2001).

Maintenance and Management

Digital libraries require constant maintenance and updating. Links to external sources have to be checked frequently. In a major system this will be done nightly, by a dedicated program. An e-mail address for contact will have to be given, and somebody will have to field the often surprisingly wide range of messages received from the general public. Major pointer systems will offer a question answering service, such as that maintained by the Internet Public Library (http://www.ipl.org/ref/QUE). This requires a major investment of time and intellectual effort. Digital library interfaces may be modern now, but will eventually become tired, old-fashioned, and in need of updating. The successful management of digital libraries requires a rare combination of computational, organizational, bibliographic, and aesthetic skills.

Preservation

Preservation is a vital element of digital library activities and touches on two distinct areas: preservation of physical originals and preservation of digital masters. Originals of photographs and broadsides will normally be retained, even after digitization. Books, which may be printed on brittle paper and may have been unbound to facilitate digitization, can be more difficult to preserve. Microfilm has been the standard answer, but libraries have recently been criticized for failing to retain the originals of microfilmed newspapers (Baker, 2001). In the future there will, no doubt, be strong pressure on libraries to preserve originals after digitization, especially when they are the last available copies.

Protection of digital masters is the second level of preservation activities. Libraries have traditionally thought in terms of centuries. Computer software only lasts a few years before being replaced by new, and often incompatible, systems. Procedures today include saving .tiff files on CD-ROMs or on digital tape, and carefully refreshing digital files at regular intervals.

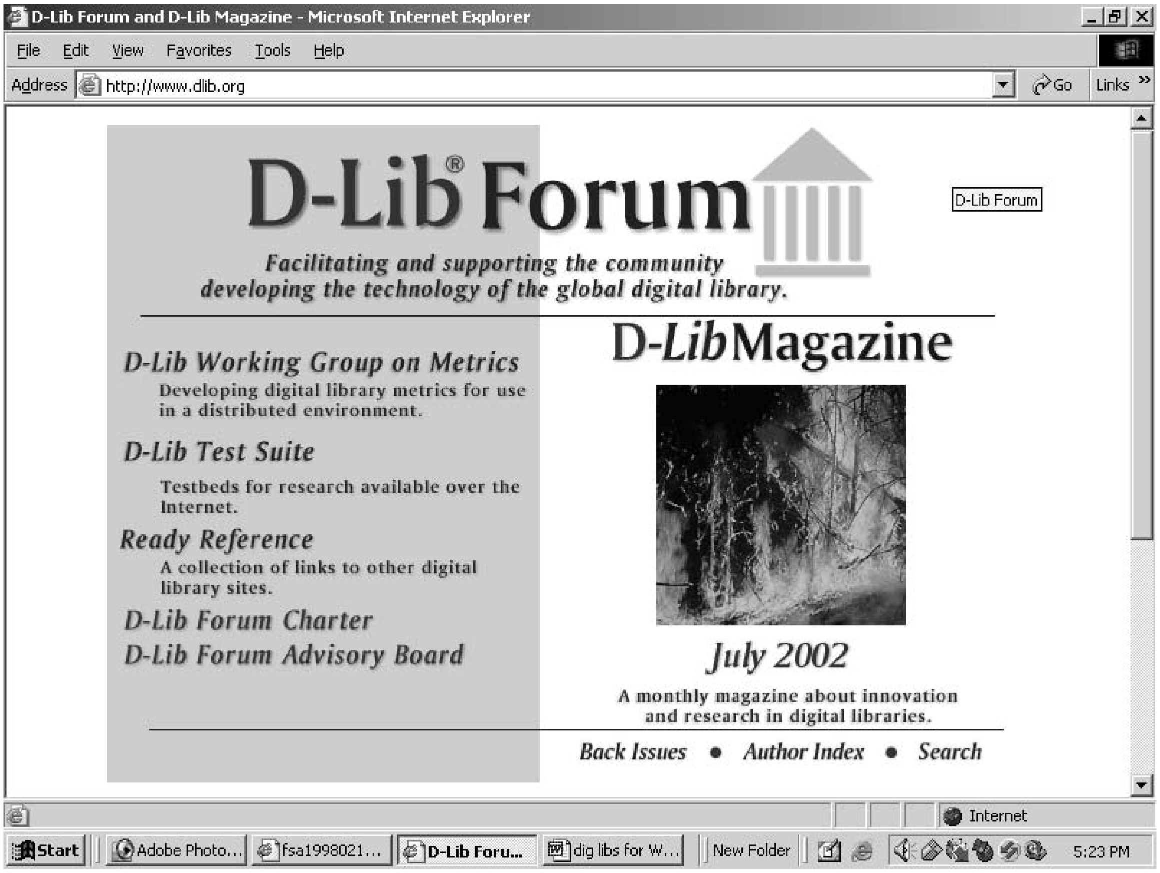
For a fascinating, publicly visible digital preservation service, see the Internet Archive and the Wayback Machine (http://www.archive.org/index.html). This permits nostalgic surfing through the Web of several years ago, and access to special collections recording specific events.

Resources in the Digital Library Field

The Digital Library Federation (http://www.diglib.org) is a Washington-based consortium of 30 major institutions. It maintains a guide to Public Access Collections at http://www.hti.umich.edu/cgi/b/bib/bib-idx?c = dlfcoll. This offers database-style access to more than 370 digital libraries. Important general information sources for the digital library field include Berkeley's SunSITE (http://sunsite.berkeley.edu/). Resource pages are maintained by Candy Schwartz at Simmons (http://web.simmons.edu/~schwartz/dl.htm), Ben Gross at Illinois (http://www.canis.uiuc.edu/~bgross/dl/), and Tom Kochtanek and Rafee Kassim at Missouri (http://whistletest.coe.missouri.edu/~rafee/iDLR/index.php). Chowdhury (2000) compares the information retrieval features of 20 important digital libraries. Major textbooks for the field have been authored by Lesk (1997) and Arms (2000). For an excellent recent printed overview of the field, see Fox (2002).

The digital library field is often financed by special government grants. Of vital importance to digital library development in the United States was the National Science Foundation's Digital Libraries Initiative. Phase 1 projects, 1994–1998, are at http://www.dli2.nsf.gov/dlione/ and were discussed by Fox (1999). Phase 2 projects can be seen at http://www.dli2.nsf.gov/ and were summarized by Lesk (1999). An important current source of finance for digitization projects in the United States is the Institute of Museum and Library Services (IMLS; http://www.imls.gov/index.htm). For examples of IMLS support, see http://www.imls.gov/closer/cls_po.asp. Research in the digital library field has been reviewed by Chowdhury (1999).

As might be expected, the digital library field is well served by e-journals. The principal U.S. e-journal is *D-lib Magazine*, published by the D-Lib Forum (Figure 17; http://www.dlib.org/).


D-Lib Forum and D-Lib Magazine - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://www.dlib.org
Go
Links
D-Lib Forum
Internet
Start
Adobe Photo...
fsa1998021...
D-Lib Foru...
dig libs for W...
New Folder
5:23 PM

D-Lib® Forum

Facilitating and supporting the community developing the technology of the global digital library.

D-Lib Working Group on Metrics
Developing digital library metrics for use in a distributed environment.

D-Lib Test Suite
Testbeds for research available over the Internet.

Ready Reference
A collection of links to other digital library sites.

D-Lib Forum Charter

D-Lib Forum Advisory Board

D-*Lib*Magazine

July 2002

A monthly magazine about innovation and research in digital libraries.

Back Issues ● *Author Index* ● *Search*

Figure 17: D-Lib Magazine is the principal United States e-journal for the digital library field.

From Britain come *Ariadne* (http://www.ariadne.ac.uk/) and the *Journal of Digital Information* (JoDI; http://jodi.ecs.soton.ac.uk/). The major U.S. conference in this area is the Joint Conference on Digital Libraries (JCDL), sponsored by ACM and the Institute of Electrical and Electronics Engineers (IEEE). The European equivalent is the European Conference on Research and Advanced Technology for Digital Libraries (ECDL).

CONCLUSION

Standards

We are still in the early stages of digital library development. Different systems are being used. Procedures for digitization, subject analysis, and indexing are emerging and will, no doubt, become standardized in the near future. Cross-collection indexing is especially important in increasing access to digital texts.

Copyright

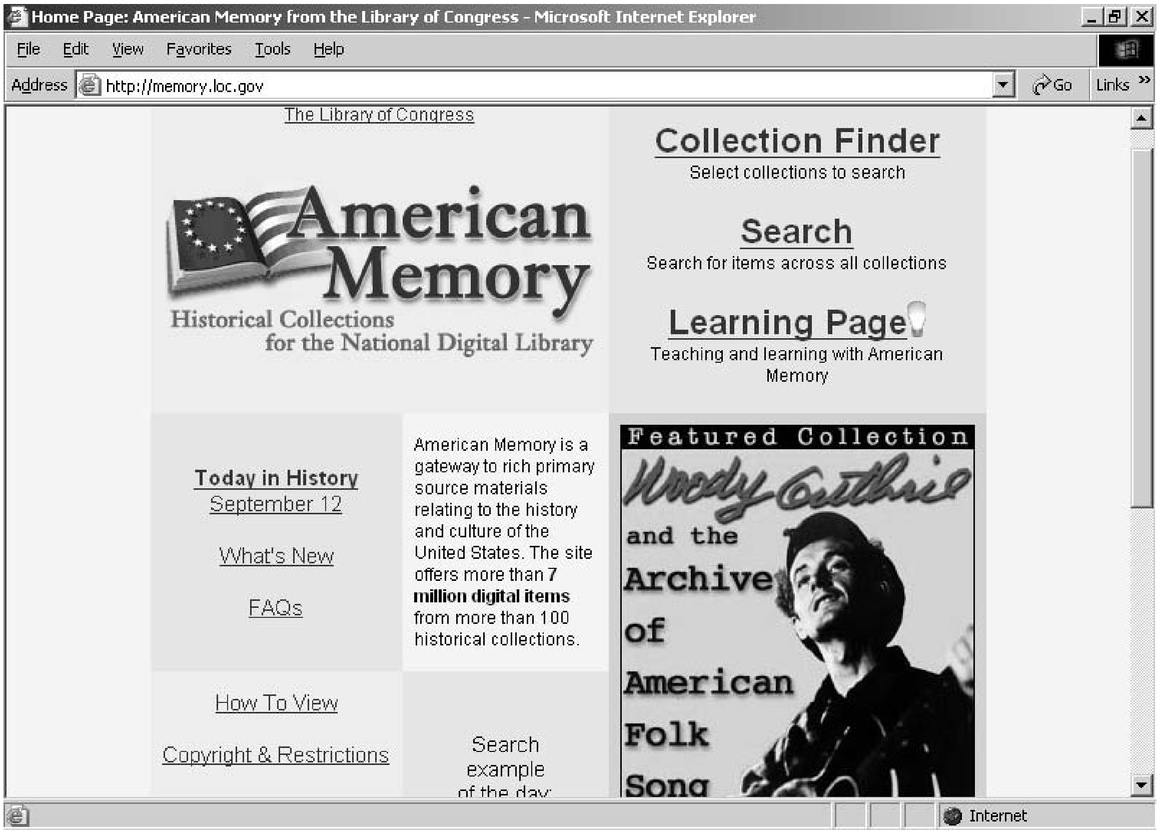
Copyright laws are a major constraint on the growth of digital libraries, because they automatically make many decades of intellectual effort unavailable for digitization. In the near term, content creators benefit from copyright protection, but the situation is confused by publishing contracts and long-term extension of copyrights. Publishers benefit from a few profitable items, but much more material is lost without significant gain to anyone. In a digital world, materials that are not available via the Internet will, in effect, almost cease to exist. They fall into the same category as oral tradition in a print society. It is also difficult to forecast how the copyright laws might be changed when publishers look on digital rights as a guarantee of future profitability.

Distance Education

More attention is being paid to distance education, especially for specialized graduate level courses. But it is difficult to see how distance education programs can be successful unless supported by adequate access to digital information.

Reading Devices

Digital development is now in a curious intermediate situation. Digital media offer clear advantages for storage and wide-area delivery of text that can be manipulated on a computer. Readers, however, still prefer the traditional end-user interface, the printed book, rather than computer screens. Early electronic reading appliances have not been successful with consumers. It is probable that the solution will shortly emerge in the convergence of the laptop computer with personal digital assistants (PDAs) and mobile telephony. This appliance would be even more attractive to the public if it were to include a digital camera, a digital recorder, and a GPS. This would create a highly portable, multifunction, use-anywhere communication and support device, which might rapidly become


Home Page: American Memory from the Library of Congress - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://memory.loc.gov
Go Links
The Library of Congress
American Memory
Historical Collections for the National Digital Library
Collection Finder
Select collections to search
Search
Search for items across all collections
Learning Page
Teaching and learning with American Memory
Today in History
September 12
What's New
FAQs
How To View
Copyright & Restrictions
Search example of the day:
Featured Collection
Woody Guthrie and the Archive of American Folk Song
Internet

American Memory is a gateway to rich primary source materials relating to the history and culture of the United States. The site offers more than **7 million digital items** from more than 100 historical collections.

Figure 18: The American Memory Historical Collections, maintained by the Library of Congress, contain more than 7 million digital items in over 100 collections.

the preferred interface for digital library consultation. Lynch (2001) points out that the new technologies have a potentially steep social price, because they provide new levels of control, monitoring, and usage restrictions for digital books. The paradox is that unless publishers are able to establish what they consider adequate levels of control, in-print books may not move to digital form. They certainly will not migrate quickly or in large numbers. It is likely that publishers will favor appliance book readers, but it is also possible that print-on-demand publishing (POD) will occupy a significant niche in the future.

The Future of Books

It is not helpful to speak of "books" in general in this context. Periodicals have undergone major changes recently, but various categories were affected differently. Weekly illustrated magazines, such as *Life*, were unable to compete with television. Abstract journals were swallowed up by online databases. Bulletins of professional associations were made redundant by discussion lists. Mimeographed poetry migrated to the Internet. But mass market magazines seem secure and local newspapers receive strong support from local advertising. A similar patchwork can be expected in the book world. Reference books, such as encyclopedias, have already largely gone into digital form. Academic publishing may well transfer to a digital environment. At least two sharply different areas should survive. Mass market paperbacks should continue to serve environments with limited access to digital libraries. Quality hand printing will continue to produce beautiful, but attainable and collectible, books. It would be wrong to say that the physical book will disappear in the digital environment. Media such as theater and wood engraving have not disappeared, although the rise of alternatives has confined them to a more limited strata of the population.

The Future of Digital Libraries

Kurzweil (1999) makes wide-ranging prophecies for the penetration of society by computers. He forecasts that by 2019, most useful 20th century documents will have been digitized; paper documents will rarely be used. The problems of digitization seem too massive for such a rapid change. The hybrid library, simultaneously storing hardcopy books and offering digital access to specialized materials, seems a more probable model for the medium term.

The long-term perspectives for digital libraries seem excellent. Libraries will, no doubt, continue to make quality and older materials available. If publishers are able to develop secure business models that guarantee significant commercial returns from digital information, they will, no doubt, migrate in force to the digital environment. Digital libraries have made a solid start, have been widely accepted, and can be expected to continue to contribute to society.

GLOSSARY

American Memory The Historical Collections of the National Digital Library, maintained by the Library of Congress (Figure 18; http://memory.loc.gov/); one of the world's largest digital libraries, with over 100 collections, a total of seven million digital items, including photographs, manuscripts, rare books, maps, sound recordings, and moving pictures.

Digital library A system that permits, via the Internet and the WWW, easy access to a collection of high-value, quality digital content, which has been selected and organized to facilitate use and is supported by appropriate services. The digital content may reflect the traditional textual orientation of many libraries or take advantage of the WWW's facilities to deliver graphics and multimedia. The term digital library is also used to refer to quality digital information and referral services whose organization reflects the traditional structure of library services.

Digitization Process of converting analog documents (e.g., paper, film, maps) to digital form.

Electronic journals (e-journals) Journals, such as scientific and professional journals, available in digital form.

Electronic theses and dissertations (ETDs) Theses and dissertations available in digital form.

Indexing The process of systematically generating pointers or keys to the content of texts, permitting readers to go directly to pages that contain specific words or deal with specific topics.

Internet Public Library Major library-style information source (http://www.ipl.org) originally set up in 1995 as a class project at the School of Information of the University of Michigan.

Metadata Data about data. Keys, guides, or indications of the content of a document.

Multimedia digital library Digital library that delivers content in a variety of formats: text, photographic images, audio, short films, and so forth.

Pointer or signpost sites Sites that do not offer their own content but direct users to Internet resources. These are considered digital libraries when organized in a manner similar to libraries.

Project Gutenberg One of the longest running digital libraries (http://promo.net/pg/or http://gutenberg.net/), founded in 1971 by Michael Hart. It has always concentrated on simple, copyright-free texts that can be read on almost any computing platform. Now offers more than 4,000 texts and is supported by a network of mirror sites.

Signpost sites See Pointer or signpost sites.

CROSS REFERENCES

See *Databases on the Web; Internet Literacy; Library Management; Online News Services (Online Journalism); Research on the Internet.*

REFERENCES

Arms, C. R. (1996). *Historical collections for the National Digital Library: Lessons and challenges at the Library of Congress*. Retrieved February 10, 2003, from http://www.dlib.org/dlib/april96/loc/04c-arms.html and http://www.dlib.org/dlib/may96/loc/05c-arms.html

Arms, W. Y. (2000). *Digital libraries.* Cambridge, MA: MIT Press.

Association of Research Libraries (2000). *Directory of Scholarly Electronic Journals and Academic Discussion Lists*. Washington, DC: The Association of Research Libraries.

Baker, N. (2001). *Double fold: Libraries and the assault on paper*. New York: Random House.

Biblio Tech Review (2001). *Z39.50, Part 1: An overview*. Retrieved February 10, 2003, from http://www.biblio-tech.com/html/ z39_50.html

Brief of Amici Curiae: The Internet archive: Filed on behalf of petitioners. (2002). Retrieved February 10, 2003, from http://www.arl.org/info/frn/copy/ia_brief.html

Chowdhury, G. G., & Chowdhury, S. (1999). Digital library research: Major issues and trends. *Journal of Documentation*, *55*(4), 409–448.

Chowdhury, G.G., & Chowdhury, S. (2000). An overview of the information retrieval features of twenty digital libraries. *Program*, *34*(4), 341–373.

Digital Library Federation (2001). *A new approach to finding research materials on the Web*. Retrieved February 10, 2003, from http://www.diglib.org/architectures/vision.htm

Emmott, B. (1998) Bad news for trees. *The Economist*, *349*(8099), 123–126.

Foster, A. L. (2002). Second thoughts on 'bundled' e-journals. *Chronicle of Higher Education*, *49*(4), A31.

Fox, E. (1999). *Digital libraries initiative (DLI). Projects 1994–1999*. Retrieved February 10, 2003, from http://www.asis.org/Bulletin/Oct-99/fox.html

Fox, E., & Urs, S. (2002). Digital libraries. *Annual Review of Information Science and Technology*, *36*, 503–589.

Kurzweil, R. (1999). *The age of spiritual machines: When computers exceed human intelligence*. New York: Penguin.

Lesk, M. (1997). *Practical digital libraries: Books, bytes, and bucks (The Morgan Kaufmann series in multimedia information and systems)*. San Francisco: Morgan Kaufmann Publishers.

Lesk, M. (1999). *Perspectives on DLI-2—Growing the field*. Retrieved February 10, 2003, from http://www.dlib.org/dlib/july99/07lesk.html

Love, C., & Feather, J. (1998). Special collections on the World Wide Web: A survey and evaluation. *Journal of Librarianship and Information Science*, *30*(4), 215–222.

Lynch, C. (2001). *The battle to define the future of the book in the digital world*. Retrieved February 10, 2003, from http://www.firstmonday.dk/issues/issue6_6/lynch/index.html

Magpantay, J.A. et al. Internet Public Library: Same metaphors, new service. (1997). *American Libraries*, *28*(2), 56–59.

Noerr, P. (2000). *The digital library toolkit* (2nd ed.). Retrieved February 10, 2003, from http://www.sun.com/products-n-solutions/edu/whitepapers/pdf/digital_library_toolkit.pdf

Stein, S. H. (U.S. District Judge) (2001). *Random House, Inc. v. Rosetta Books LLC, F. Supp. 2d, 2001 U.S. Dist. Lexis 9456 (S.D.N.Y. 2001)*. Retrieved February 10, 2003, from http://www.law.cornell.edu/copyright/cases/ebooks.htm

Saffady, W. (1995). Digital library concepts and technologies for the management of library collections: An analysis of methods and costs. *Library Technology Reports*, *31*, 221–380.

Schwartz, C. S. (2000). Digital libraries: An overview. *Journal of Academic Librarianship*, 26(6), 385–93.

University of Virginia Electronic Text Center (1998). *Text scanning: A basic helpsheet*. Retrieved February 10, 2003, from http://etext.lib.virginia.edu/helpsheets/scantext.html

Digital Signatures and Electronic Signatures

Raymond R. Panko, *University of Hawaii at Manoa*

INTRODUCTION

When we send letters, we sign them to indicate that they are from us. When we sign contracts, we are expressing our willingness to abide by the terms of the contract. We cannot later repudiate the contract because our signature binds us. Signing is also possible in the electronic world, and it generally serves the same purposes.

There are three related terms we use in this article. An *electronic signature* (e-signature) is any signing method that is used with computers and networks. It is the broadest concept. It includes such things as clicking a button to indicate that we accept the terms of a program's end-user licensing agreement.

More narrowly, there are two general ways to add signature blocks to outgoing messages. Digital signatures are signature blocks created with public key encryption. Message authentication codes (MACs) also are per-message signature blocks, but they are created using hashing. MACs are also called key-hashed message authentication codes (HMACs).

In our discussion, we begin with the narrowest and most familiar technology, digital signatures. We then discuss MACs and, finally, electronic signatures broadly.

BACKGROUND

Applicant, Verifier, and True Party

A prime reason for electronic signing is authentication. In authentication, there are two main parties. The *verifier* wishes to determine the identity of the *applicant*—the party wishing to have his or her identity authenticated. Applicants are sometimes called supplicants.

In addition, the *true party* is the person the applicant claims to be. (The applicant may be an impostor.) The person who signs the document is the *signatory*; this may be the true party or someone authorized by the true party to sign for the true party.

Key-Based Authentication

As discussed in the chapter on authentication, authentication can be based on something the person knows (such as a reusable password), something a person is (biometrics), or some other distinguishing characteristic.

Digital signatures and MACs are based on the applicant knowing a secret key. Digital signatures are based on public/private encryption key pairs and require the applicant to know the true party's private key. MACs require the person wishing to be authenticated to know the symmetric key the true party shares with the verifier.

Threat Model

In normal authentication, the biggest danger is that the *applicant* is an impostor who tries to impersonate the true party in one or more transactions. This danger also is a key element in electronic signature threat models.

In addition, there is a danger that the true party will later falsely repudiate messages and contracts that he or she signed electronically, claiming that these were signed by an impostor. Against this threat, we would like to have nonrepudiation, that is, the ability to provide proof that the true party actually did sign the messages or contracts.

In simple authentication, the *verifier* is assumed to be the "good guy." However, electronic signatures should also protect the true party from verifier malfeasance. For instance, the verifier might fabricate a message or contract, add a false signature, and then claim that the true party sent the signed message or contract.

While the true party and verifier are communicating, an *attacker in the middle* may insert a single fabricated message into an ongoing dialog. Or an attacker in the middle might delete a message or simply replay an earlier message. Initial authentication at the start of a dialog will not protect against such attacks.

DIGITAL SIGNATURES

Digital signatures are used in message-by-message authentication. A digital signature is a block of bits attached to each outgoing message to prove the sender's identity. This greatly reduces attacker-in-the-middle threats. t also provides nonrepudiation on a message-by-message basis. Figure 1 illustrates the process of creating and verifying digital signatures.

Creating the Digital Signature

The sender creates a message to be sent. In cryptographic terminology, this is the plaintext or original plaintext. The name is a bit misleading because the message may not be limited to text, but for historical reasons, the term plaintext remains in use.

The sender/applicant will have to sign something (encrypt it with his or her private key) for authentication to be possible. However, public key encryption is very processing-intensive, so it can only be used on small blocks of bits—not on large messages.

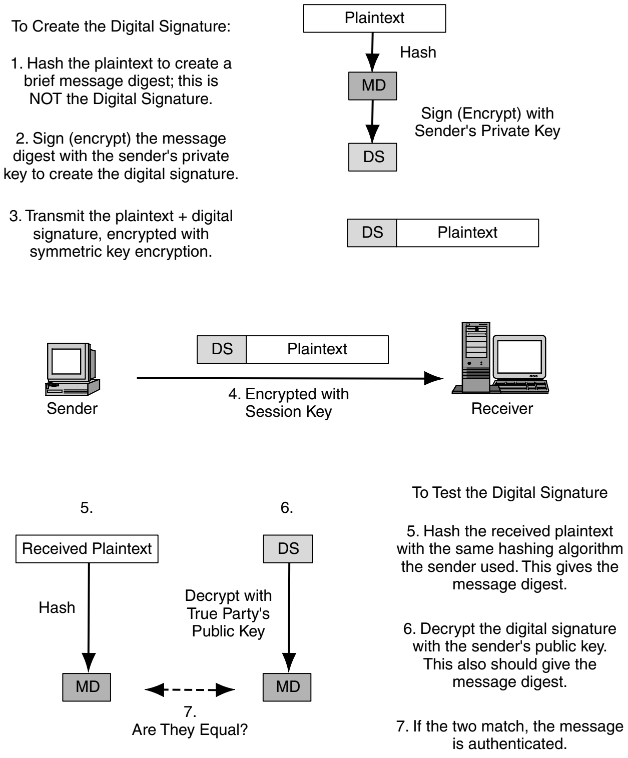

To Create the Digital Signature:
1. Hash the plaintext to create a brief message digest; this is NOT the Digital Signature.
2. Sign (encrypt) the message digest with the sender's private key to create the digital signature.
3. Transmit the plaintext + digital signature, encrypted with symmetric key encryption.
Plaintext
Hash
MD
Sign (Encrypt) with Sender's Private Key
DS
DS
Plaintext
DS
Plaintext
Sender
4. Encrypted with Session Key
Receiver
To Test the Digital Signature
5. Hash the received plaintext with the same hashing algorithm the sender used. This gives the message digest.
6. Decrypt the digital signature with the sender's public key. This also should give the message digest.
7. If the two match, the message is authenticated.
5.
6.
Received Plaintext
DS
Hash
Decrypt with True Party's Public Key
MD
MD
7.
Are They Equal?

Figure 1: Digital signature creation, transmission, and verification.

To create something small to sign, the sender's software first hashes the original plaintext message. Hashing is a mathematical process that can be applied to a string of bits of any length and that will produce a result (called a hash) that has the same short length no matter how long the input string is. For instance, the MD5 hashing algorithm always produces a hash of 128 bits, whereas the SHA-1 hashing algorithm always produces a hash of 160 bits. So message digests will be either 128 bits or 160 bits, depending on which of these hashing algorithms is used.

The hash of the original plaintext is called the message digest. The message digest is not the digital signature itself but rather the basis for creating the digital signature.

The applicant/sender wishes to authenticate himself or herself using something only the true party should know. This is the true party's private key. If someone is given a public key–private key pair, he or she should guard the private key jealously. However, their public key is not secret and can be shared freely.

Therefore, the applicant/sender signs the message digest with his or her private key, that is, encrypts the message digest with his or her private key. The result of this encryption is the digital signature.

Transmission and Confidentiality

After creating a digital signature, the applicant/sender creates a composite message by concatenating the bits of the digital signature to the bits of the original plaintext message. We will call this the composite message. (Terminology here is not standardized.)

Next, the applicant/sender normally encrypts the composite message with the symmetric key that he or she shares with the verifier/receiver. This provides confidentiality, meaning that no one can read the original plaintext en route. Note that this step has nothing to do with authentication, and it is possible to wish to have authentication without confidentiality. However, confidentiality is normally desired during transmission.

Symmetric key encryption is used rather than public key encryption because the composite message may be quite long. As noted earlier, public key encryption is unfeasible for long messages. Symmetric key encryption, in contrast, is efficient enough for longer messages.

The applicant/sender now transmits the composite message encrypted with symmetric key encryption to the verifier/receiver. This message cannot be read en route by an attacker in the middle.

The verifier/receiver decrypts the transmitted message using the symmetric key it shares with the applicant/sender. This restores the composite message.

Verification

Now it is time for the verifier/receiver to verify the authenticity of message. This involves recomputing the message digest in two ways and comparing the results.

One way to recompute the message digest is to rehash the original plaintext. The applicant/sender hashed the original plaintext message to create the message digest.

The verifier/receiver rehashes the original plaintext message using the same algorithm the applicant/sender used. Hashing is a repeatable process, meaning that the verifier/receiver will get the same resulting hash the applicant/sender obtained. This, of course, is the message digest.

The second way is to decrypt the digital signature. The digital signature was created by encrypting the message digest with the true party's private key. The verifier/receiver, in turn, decrypts the digital signature with the true party's public key, which is widely known. Public key encryption is reversible, so this decryption will give the message digest.

In the final step, the verifier/receiver compares the two computed message digests. If they are the same, then the digital signature was created with the true party's private key. If the message digest was encrypted with an impostor's private key, decrypting the digital signature with the true party's public key would not give the message digest. Only the true party would know the true party's private key, so the message is authenticated.

Benefits

Digital signatures provide three important benefits. One is message-by-message authentication. This guards against the insertion of a fabricated message in a dialog's message stream by an attacker in the middle.

A second benefit is message integrity, that is, proof that a message has not been tampered with en route. If an attacker has deliberately modified a message or if there has been a technical transmission error, the two message digests will not match. The verifier/receiver will discard the message.

The third benefit is nonrepudiation. If the message was signed by the true party, then the true party cannot disclaim responsibility for the message without arguing that he or she lost control of the private key, which itself may be considered negligence. If a private key is stolen, of course, it can be used as a rubber stamp to sign documents. However, for nonrepudiation, the verifier/receiver must keep the original composite message so that authentication can be verified in court or by an expert.

The Problem of the True Party's Public Key and Digital Certificates

Digital signature verification requires the verifier/receiver to know the true party's public key. This seems simple, but it is fraught with danger. For instance, suppose the applicant/sender sends the verifier/receiver a public key claiming that it belongs to the true party. If the applicant/sender is an impostor, of course, he or she will send his or her own public key rather than the true party's public key. If the verifier/receiver accepts this impostor's public key as the true party's public key, the impostor will be "verified" as the sender of all messages. This is public key deception.

To guard against such deception, the verifier/receiver must get the true party's public key from a trusted third party. As discussed in the chapter on public key infrastructures (PKIs), organizations called certificate authorities (CAs) provide such information in the form of digital certificates. To use digital signature authentication, the verifier/receiver needs to get the true party's digital certificate from a trusted certificate authority.

There is a great deal of confusion about digital certificates. The main thing to keep in mind is that the essential information that digital certificates provide is the true party's name and the true party's public key. Anyone claiming to be the named true party should be able to create digital signatures that can be tested with the public key enclosed in the digital certificate.

A digital certificate generally does not vouch for the trustworthiness of the party named in the digital certificates. Although some CAs provide compensation if a named party behaves badly, few do. Vouching for trustworthiness is not what digital certificates are designed to do. Digital certificates are designed to tell you the public key of the named party.

At the same time, companies and individuals who do behave badly may have their certificates revoked before the expiration date on the digital certificate. CAs maintain certificate revocation lists (CRLs) of such digital certificates.

Of course, if the verifier/receiver contacts the certificate authority, the authority will not send the verifier/receiver a revoked certificate. However, it is crucial for the verifier/receiver who gets the digital certificate from another party, for instance the applicant/sender, to check the certificate authority's CRL to be sure that the certificate has not been canceled.

The certificate authority also has a private key and a public key. The CA adds a digital signature to every certificate it creates, signed with the CA's private key. Popular CAs have public keys that are well known, so it is easy for verifier/receivers to check any digital certificate sent to them. As noted earlier, digital signatures provide message integrity, ensuring that the digital certificate has not been modified, say by entering an impostor's public key in place of the true party's public key.

In practice, every browser today comes with the ability to read digital signatures automatically, without the user's intervention. Browsers also come with the public keys of several root certificate authorities, so they can also handle most digital certificates. What the user sees is merely a notification that a particular document came from a particular, named entity. When this notice appears, the user can feel confident that the message really did come from that person or organization.

MESSAGE AUTHENTICATION CODES (MACs)

Message authentication codes (MACs) are similar to digital signatures. Both are blocks of bits appended to original plaintext messages. However, although digital signatures are created using public key encryption, MACs are created using hashing. Figure 2 illustrates the creation, transmission, and use of message authentication codes.

Speed

The main advantage of MACs over digital signatures in processing speed. Public key encryption used in digital

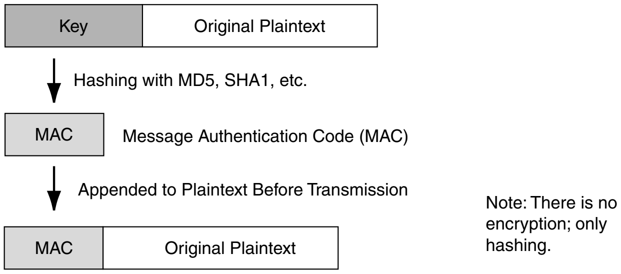

Key
Original Plaintext
Hashing with MD5, SHA1, etc.
MAC
Message Authentication Code (MAC)
Appended to Plaintext Before Transmission
MAC
Original Plaintext
Note: There is no encryption; only hashing.

Figure 2: The creation, transmission, and verification of message authentication codes (MACs).

signatures is very slow, even if only the message digest is encrypted. In contrast, the hashing used in message authentication codes requires far less processing time. Quite simply, MACs place much less of a load on the machines of both the applicant/sender and the verifier/receiver.

Symmetric Key

The applicant/sender and the verifier/receiver share a single key in hashing. Each uses this key both to create MACs and to verify MACs.

Often, the symmetric key used for MAC authentication is different from the symmetric key used for confidential transmission. In such cases, the two parties have at least two symmetric keys that they share.

Creating a Message Authentication Code

To create a MAC, the applicant/sender again begins with the message to be sent—the original plaintext. Next, the applicant/sender appends the symmetric key to be used in MAC creation to the original plaintext message.

Next, the sender hashes the combined original plaintext and symmetric key with either MD5 or SHA-1 to create a hash of 128 bits or 160 bits, respectively. This hash is the message authentication code.

Transmission

To transmit the message, the applicant/sender appends the MAC to the original plaintext message. This is the composite message that actually will be transmitted.

Next, for confidentiality, the applicant/sender normally encrypts the composite message with a different symmetric key shared by the two parties. The applicant/sender then transmits the resultant cipher text. Interceptors will not be able to decrypt the cipher text back to plaintext because interceptors will not have the shared symmetric key used for confidentiality.

The verifier/receiver undoes the encryption for confidentiality by decrypting the cipher text with the shared symmetric key used for encryption. This gives the verifier/receiver the composite message consisting of the original plaintext plus the MAC.

Verifying the MAC

Verifying the MAC is simple. The verifier/receiver takes the plaintext and appends the symmetric key used for authentication. The verifier/receiver then hashes the plaintext plus key using the same hashing algorithm the applicant/sender used. This should give the MAC transmitted with the message.

If this process successfully reproduces the MAC, then the sender must know the symmetric key used for authentication. Only the true party and the verifier/receiver should know this key. The MAC must have been created by the true party.

Benefits

Like digital signatures, MACs provide authentication. Also like digital signatures, MACs provide message integrity. If the message is altered en route either deliberately or by transmission errors, the verification process will not reproduce the MAC, and the message will be discarded.

Lack of Nonrepudiation Assurance

MACs really prove that someone who knows the symmetric authentication key created the MAC. Obviously, this could be the true party acting as the applicant/sender.

However, although it is easy to overlook the fact, the verifier/receiver could also have created the MAC because the verifier/receiver also knows the symmetric authentication key. Why would a verifier/receiver fabricate a MAC? The answer is that the verifier/receiver may be dishonest and wish to claim that the true party sent a message that the true party never sent, for instance a message agreeing to a dubious contract.

Thanks to the possibility of verifier/receiver misbehavior, MACs cannot provide nonrepudiation. A dishonest true party can repudiate a legitimate message claiming that the verifier/receiver really created it. In court cases, jurors would have to decide who to believe—hardly an easy undertaking. Where nonrepudiation is an issue, MACs are dangerous.

IPsec

One reason to consider MACs is that they are the default message-by-message authentication mechanism in IPsec (IP security) standards. These are increasingly being used in virtual private networks (VPNs) to ensure confidentiality, authentication, message integrity, and other benefits in dialogs between partners. If IPsec does become very widely used and if MACs continue to be the default authentication/integrity method, the lack of nonrepudiation could become a serious problem for business transactions.

OTHER ELECTRONIC SIGNATURE TECHNOLOGIES

Digital signatures accompanied by digital certificates are the gold standard in electronic signature technologies, and MACs are good despite their lack of nonrepudiation. However, quite a few other types of e-signature technologies are possible and may make sense in particular situations.

Typed Signatures and Scanned Physical Signatures

In the simplest e-signature methods, the sender merely types his or her name at the end of a message or includes a scanned copy of his or her written signature. Although typed signatures and scanned physical signatures are allowed under most e-signature regulations, they are not likely to stand up in court because they are so easily forged.

Click Agreements

When you purchase software, you typically are required to click on a dialog box button to show that you accept the user licensing agreement. Click agreements generally are difficult to enforce in court because of the difficulty proving who actually clicked on the button.

However, enforceability often is not the goal. Rather, click agreements often serve primarily to create a ceremony in which the person formally makes a commitment. This brings the seriousness of the situation to the person's attention. In addition, when people make explicit commitments, they may be more likely to keep them.

Authenticated Sessions

Many transactions are sessions in which the two parties exchange a long series of messages. With various mechanisms, it is possible to authenticate the user at the beginning of the session and perhaps occasionally during the session to ensure that the person is still there. Although less secure than digital signatures and MACs, which authenticate every message, authentication at the beginning of a session (initial authentication) provides some assurance that a certain party is sending the messages.

Most initial authentication systems rely on reusable passwords, which people (or software processes) use each time they log in for a certain period of time.

The problems with reusable passwords are well known. Unless the system enforces strong passwords, people tend to use easily guessed passwords that can be cracked in a few seconds by a password-cracking program. People also tend to write their passwords somewhere, often on their computer monitors.

A more subtle problem is lost passwords. About a quarter of all calls to help desks are for lost passwords. The help desk operator can perform a "password reset," giving the account a new password. However, there is danger in giving out for new passwords over the telephone. The caller may be an imposter, not the real account holder. Although some password reset systems require the caller to answer questions that only the true account holder should know, most of these questions are easily guessed if the impostor has done his or her research.

Another means of authentication is access cards and tokens. If you have stayed at a hotel room recently, you probably were given an access card that allowed you into your room. Most such access cards have magnetic stripes containing information that allows access to the room; smart card versions have microprocessors and memory for more sophisticated identity checking.

Even if you get a key, the key is likely to be a physical token that contains access information. Plugging it into the door allows the door reader to query a central system for access permission. There also are tokens that plug into the USB ports of personal computers for access to those machines.

Another type of token requires the user to enter a PIN number on a small (generally) numerical keypad. The token then shows a temporary password on its display. The user must use this temporary password to log into a computer system.

Access cards and tokens provide good security, but they are easily lost or stolen. There must be a quick way to disable lost access devices as well as a way to reissue them in ways that impostors cannot exploit.

Another approach to handling authenticated access sessions at Web sites is visit traces, recording the paths users take through them, including click agreements they have made. Documentation that a person saw certain information can be convincing evidence in court and may prompt a person to drop repudiation claims. However, visit trace logs must be secured against tampering by the logkeeper if they are to be useful in court.

Biometrics

One form of authenticated session technology is biometrics. It is new and complex, so we will give it its own section.

"Biometrics" comes from the words "bio," meaning biological life, and "metrics," meaning measurement. Some types of biometrics measure bodily dimensions, such as fingerprints, iris patterns in the eye, and facial features. Other types measure activities, such as motions and pressures involved when in signing a name or the temporal patterns of password typing.

The advantage of biometric authentication is that it does not require the applicant to carry something that can be lost and (usually) does not require the applicant to remember anything. Despite jokes to the contrary, we never actually forget our heads at home. A major hope is that biometrics will replace reusable passwords as the dominant form of authentication.

A major problem with biometrics, however, is that there are serious disagreements over the error rates involved in biometric measurements. Many vendors make impressive claims about accuracy, but these often are based on tests conducted under ideal conditions and may not be representative of accuracy in the real world.

There are two basic types of errors in biometrics, indeed in all access control methods. False acceptance rates (FARs) verify or identify someone who should be rejected. High FARs means impostors are getting in and forming a strong basis for repudiation. A failure to test FARs may make access data difficult to defend in court.

At the opposite end of the spectrum is false rejection rates (FRRs). FRRs tell you what percentage of legitimate applications are rejected. Although FRRs do not harm security, significant FRRs may make a system unacceptable to users. Many systems allow applicants to attempt to authenticate themselves several times to reduce FRRs.

Another issue in biometrics is user acceptability. For instance, some users may refuse to use a fingerprint system because of its criminal connotations. Others will reject systems they fear may harm them; for instance, some people believe that eye identification systems shoot laser beams into their eyes. Still others reject systems that are difficult to use, such as iris scanning systems, which require proper eye placement. In general, if a significant number of users refuse to use the system, the loss in revenues and other values may make the system completely cost ineffective.

Another problem with biometrics is that the different technologies vary widely in cost and accuracy. Selecting a biometric method is an important task. Not surprisingly, the most expensive methods tend to be the most accurate. In addition, the cost of biometric readers and other system components may be prohibitive.

A final problem is that we do not yet have comprehensive standards for biometrics. As a consequence, choosing biometric authentication today generally means getting locked into a single vendor.

None of the Above

The last alternative in e-signatures is simply not to use them. In many cases, the benefits may not be worth the costs and damaged user relationships.

SELECTING AN ELECTRONIC SIGNATURE METHOD

Selecting an electronic signature method is not a technical decision. It is a business decision. Like any business decision, it requires the selector to understand the business situation before considering anything else.

Some e-signature systems serve closed communities, such as individual corporations or consortia of firms. Others serve open communities, such as a vendor and its customers. In open communities, it is difficult to impose stringent e-signature requirements. For instance, in consumer e-commerce, it is traditional in the SSL/TLS encryption methodology that is used in almost all transactions to require the merchant but not the consumer to have a digital signature and digital certificate. Note that e-signature implementation can be asymmetrical, with different requirements imposed on the two sides. In addition, in an open community, the general lack of e-signature and PKI standards and the need to coordinate rollouts in many firms tends to require long lead times.

There are two forms of authentication, verification and identification. In verification, the person claims to be a particular person, for instance by typing in an account name. The authentication system then only has to see whether the password typed or the other authentication data given is correct for the account. If so, the applicant is verified as the true account holder. In identification, the applicant does not claim a particular identity. The applicant provides authentication data, and the identification system matches that data against that of *all* the accounts in the identification database. If the best match that is selected meets closeness-of-fit criteria, the applicant is identified as that person (or software process) in the database. Identification is more difficult than verification and thus has higher error rates. It also may lack the intentionality that is normally present in signing activities and so may not be enforceable in court.

Security Requirements

Signatures of any kind exist to validate agreements. To be effective, they must be safe from use by impostors and from other security threats.

In security, one must always consider threat severity, which is the likely cost of a security incursion times the probability an incursion will take place. It is, in other words, the expected value of loss from a threat. If the threat severity is low and is likely to remain smaller than the cost of implementing an e-signature system, then implementing the system will not make economical sense.

In general, e-signatures should be viewed as security techniques, and security always must consider risk management—the balancing of risks and countermeasure costs. Generally speaking, however, the more expensive the transaction, the more expensive the e-signature.

Key length is an important security concern when private keys are used to sign documents or message digests. Documents that must be kept secure for many years must be signed with longer keys than documents whose signatures only have to be verifiable for a few years. However, longer keys mean longer processing times and therefore higher costs. The period of sensitivity is a crucial determinant of key length.

Legal Goals

Another consideration in selecting an electronic signature technology is a firm's legal goals. Many countries require that contracts worth over a certain amount of money be signed to be valid. In U.S. commercial law, for instance, contracts over $500 or lasting a year or more must be signed to be valid. If the goal is to meet this requirement, even the least-secure e-signing methods may be acceptable.

As noted earlier in this chapter, one consideration is whether or not to create a ceremony of commitment, in which a person must explicitly, through an action, acknowledge ownership of a document. Again, if this is the main goal, even nonsecure e-signing methods may be acceptable.

If the goal is nonrepudiation, then a stronger e-signature method is needed. The only method available today that provides strong technical nonrepudiability is use of digital signatures, which require a PKI for digital certificates. As noted earlier, it is necessary to keep the ciphertext version of the digital signature so that the decryption can be tested in court.

Although technical nonrepudiability is good, juries are likely to decide contested cases. If the complainant is more believable than the defendant, the jury may disregard the defense's argument that the e-signature method used does not provide technical nonrepudiation. If the defendant is more believable, the jury may side with the defense even if a digital signature and digital certificate are used.

Suitability

Another consideration in selecting an e-signature methodology is whether the system can be implemented. One question to ask when considering the various methods is what, in any given firm, is technically feasible. Some choices may not be feasible, and some that are may be outside the firm's resources to implement it. PKIs are especially problematic.

Of course, the firm must consider the cost of implementing a system, including the cost of the technology itself and the cost of installing it. In addition, electronic signature processing slows computer processing often enough to require upgrading to faster hardware.

The last major consideration in determining the suitability of a method is whether users will accept it. As noted earlier, fingerprints, iris scanning, and other techniques may offend users or make them uncomfortable. The company must also consider the cost of lost business if some users refuse to use e-signatures and therefore stop doing business with the firm. On the other hand, a firm may gain revenues if it develops a reputation for having strong security in general.

Obviously, individual home consumers are unlikely to be willing to get digital certificates unless they see strong benefits from doing so. As a result, most firms are unlikely to require consumers to use digital certificates for fear of losing business.

LEGAL AND REGULATORY ENVIRONMENT

One consideration that is especially difficult to discuss cleanly is the legal and regulatory environment of electronic signatures. Laws vary around the world and even within countries. Few of these laws, furthermore, have been tested in court. Regulation of certificate authorities and other aspects of electronic signatures barely exists, even in countries that have begun to back legislation with regulation.

In the United States, a number of states created electronic signature laws before the federal government created the Electronic Signatures in Global and National Commerce Act, better known as E-SIGN, in 2000. As the name states, E-SIGN governs only national (interstate) commerce. No court cases have yet appeared to test this area.

Quite a few states have created their own electronic signatures laws. In 1999, before Congress acted, the National Conference of Commissioners on Uniform State Laws adopted the Uniform Electronic Transactions Act (UETA), and a number of states based their laws on this act. However, past court rulings have found that many intrastate activities affect interstate commerce and so are governed by federal laws.

In 1999, the European Parliament and the Council of the European Union created Directive 1999/93/EC, On a Community Framework for Electronic Signatures. This directive did not create laws but rather directed the member countries to create e-signature regulation. The first country to do so was Ireland, in July 2000. Germany followed in May 2001.

Elements

The simplest element in electronic signature laws is whether legal validity has been established. As noted earlier, most jurisdictions require that contracts worth more than certain amount of money be signed to be valid. Almost all e-signature laws provide the weak protection of saying that contracts cannot be invalid simply because they are signed electronically. This does not ensure that a particular e-signature methodology will stand up in court if one side repudiates the document.

One consideration in selecting an electronic signature technology is whether a country's e-signature law allows all types or only some types of electronic signatures. The U.S. E-SIGN Act is intentionally vague, not mentioning specific e-signature technologies. The EC Directive mentions several specific technologies but does not limit itself to them.

The EC Directive defines an electronic signature in general as "data in electronic form which are attached to or logically associated with other electronic data and which serve as a method of authentication." The EC Directive also specifies *advanced e-signatures,* which basically are digital signatures based on qualified certificates—the strongest type of electronic signature. These advanced signatures are given a certain degree of privileged status, which is reasonable because of their strength. Most important, advanced electronic signatures are viewed as equivalent to hand signatures in legal proceedings.

Another consideration is whether e-signatures are permitted for all types of documents or only some. For instance, in the United States, the federal E-SIGN law forbids certain documents to be signed electronically because of potential harm to consumers. These exclusions were not present in early versions of the laws, which failed to pass because of insufficient consumer protection. Among the documents excluded are wills, trusts, adoptions, divorces, other family court matters, court documents, utilities cancellation notices, notices of foreclosures, eviction notices, insurance cancellations, warnings about the transportation of hazardous materials, and the repossession of primary residences.

Another key consideration in selecting an electronic signature technology is whether one side can force the other side to accept electronic signatures or whether using e-signatures is voluntary. Early versions of the U.S. E-SIGN law were rejected by Congress because they failed to say explicitly that people cannot be forced to accept electronically signed documents. The final E-SIGN Act says that consumers must be notified of options, including the mandatory option of paper-only transactions, must give consent, and must demonstrate the ability to store and access digital documents.

Digital signatures require digital certificates from certificate authorities. A major issue is whether a country will regulate certificate authorities. The U.S. E-SIGN Act leaves everything to industry. The EC Directive, in contrast, specifies that each country should establish a regulatory framework within the country for certificate authorities and related services, such as registration services, time stamping services, and directory services (Article 1a).

CONCLUSION

Documents of importance have always had to be signed, whether by hand or with a signing device, such as a signet ring or a Japanese hanko. Electronic signatures merely extend this long-standing practice to the electronic world.

Although electronic signatures are increasingly recognized as legitimate, the choice of which electronic signature technology to use is complex. Using a digital signatures accompanied by a digital certificate is the legally strongest valid way to sign a message, but the public key infrastructures that support digital certificates are not well established or standardized. Consequently, many firms will turn, at least initially, to less rigorous forms of electronic signing, which are easier to implement but cannot provide strong guarantees.

In addition, although technical considerations are important, many cases end up in court, and juries are not always swayed by technical arguments.

GLOSSARY

Access cards Have magnetic stripes containing information that allows you into, for instance, a hotel room. Smart card versions have microprocessors and memory for more sophisticated identity checking.

Advanced e-signatures In the European Union electronic signature directive, digital signatures based on qualified certificates.

Applicant The party wishing to have his or her identity authenticated.

Attacker in the middle Party who may insert a single fabricated message into an ongoing dialog, delete a message, or replay an earlier message.

Authentication data Where an identification system matches user-supplied data against all accounts in the identification database to determine who the user is.

Biometric authentication The authentication of a person based on body measurements.

Ceremony of commitment Act of signing a document that results in heightened awareness of the gravity of the situation on the part of the signer.

Certificate authorities (CAs) Organizations that create and distribute digital certificates.

Certificate Revocation List (CRL) List of a certificate authority's certificates that have been revoked before the termination date listed on the certificate.

Click agreements Electronic signatures created by the user clicking on a button that says, for example, "I agree."

Digital certificates Documents that give a named party's public key and other information.

Digital signatures Signature blocks created with public key encryption. They are created by encrypting message digests with the applicant's private key.

Directive 1999/93/EC European Union electronic signature directive to the EU member nations.

Electronic signature (e-signature) Any signing method that is used with computers and networks.

E-SIGN U.S. electronic signature law.

False acceptance rates (FARs) Percentage of times a person is authenticated when he or she should not be.

False rejection rates (FRRs) Percentage of times a person is not authenticated when he or she should be.

Hashing A mathematical process that can be applied to a string of bits of any length and that will produce a result (called a hash) that has the same short length no matter how long the input string is.

Identification Process in which the applicant does not claim a particular identity. When the applicant provides authentication data, the identification system matches the data against all the accounts in the identification database to determine who the applicant is.

Message authentication codes (MACs) Per-message signature blocks, created using symmetric key encryption. MACs are also called key-hashed message authentication codes (HMACs).

Message digest The result of hashing a plaintext message. This is the first step in creating a digital signature.

Message integrity Proof that a message has not been tampered with en route to its destination.

Nonrepudiation When the sender cannot plausibly claim that a message did not really come from him.

Password reset Gives an account a new password when a reusable password is forgotten.

Period of sensitivity Period of time during which a file must be kept confidential.

Public key deception Process in which an impostor sends his own public key, claiming that it is the true party's public key.

Repudiate When a party claims that she did not send a message apparently sent from it.

Reusable passwords Passwords used repeatedly. Most passwords are reusable passwords.

Scanned physical signatures Electronic signatures formed by scanning a written signature and inserting the image in the document as a signature block.

Signatory Party who actually signs the document. May be the true party or someone or something to which the true party delegates signing authority.

Supplicant Another name for an applicant.

Tokens Physical devices used to authenticate a person.

True party The person or object the applicant claims to be.

Typed signatures Electronic signatures in which the sender merely types his or her name.

Validity Characteristic of a document that allows it to be presented as evidence in a court.

Verification When the person claims to be a particular person, for instance when typing in an account name.

Verifier The party wishing to determine the identity of the applicant.

Visit trace A list of the locations a user has visited at a Web site.

CROSS REFERENCES

See *Authentication; Biometric Authentication; Digital Identity; Electronic Commerce and Electronic Business; Encryption; Internet Security Standards; Passwords; Secure Electronic Transmissions (SET); Secure Sockets Layer (SSL).*

FURTHER READING

Digital signature guidelines: Legal infrastructure for certification authorities and electronic commerce. (1996). Chicago, IL: American Bar Association.

Ford, W. (2000). *Secure electronic commerce: Building the infrastructure for digital signatures and encryption.* (2nd Ed.). Prentice-Hall PTR.

Hammond, B. (2002). *Digital signatures.* McGraw-Hill Osborne Media.

Panko, R. (2004). *Business computer and network security.* Englewood Cliffs, NJ: Prentice-Hall.

Piper, F., Blake-Wilson, S., & Mitchell, J. (2000). *Digital signatures: Security and control.* Information Systems Audit and Control Foundation.

Pfitzmann, B. (1996). *Digital signature schemes: General framework and fall-stop signatures.* Berlin: Springer-Verlag.

Disaster Recovery Planning

Marco Cremonini, *Università di Milano, Italy*
Pierangela Samarati, *Università di Milano, Italy*

INTRODUCTION

For many years, corporations have placed most of their critical information assets into automated systems and, more recently, they have adopted business models largely based on the Internet. While allowing for more efficiency, cost reduction, and competitive advantage, these developments have also increased the dependency on the information technology (IT) infrastructure and therefore of the potential economic losses caused by its failure. More than ever today a loss of access to the IT infrastructure will prove to be in many cases fatal for the business of affected companies. It is therefore crucial for companies to consider precise plans for counteracting possible events that may cause such interruptions. Examples of companies that failed to recover their businesses after a prolonged interruption of IT functions, due to floods or fires, for example, are well known and documented (Hiles, 2002; Kaye, 2001; Noakes-Fry & Diamond, 2001; Toigo, 1999, 2001a).

The disaster recovery planning process consists of a set of activities aimed at reducing the *likelihood* and the *impact of disasters* on critical business assets (NIST, 2002; Toigo, 1999; Wold, 2002). Note that, while the term disaster recovery has been traditionally associated with protecting against external catastrophic events (e.g., floods, fires, earthquakes), we adopt in this chapter a broader definition of disaster recovery, which includes also business continuity planning, and therefore the consideration of failures to technological equipments and software.

E-business and Internet connections are shortening recovery-time requirements and changing the way we think about disaster recovery planning: full 24 × 7 operation and "zero downtime" are now commonly seen as a business necessity. While traditional countermeasures to face natural disasters should still be provided, a wide spectrum of other risks (e.g., cyberattacks) must be considered. Also, critical business processes must be analyzed and recovered much faster than in the past: downtimes of days (the usual time window for natural disasters) are definitely no longer sustainable for Internet-based businesses.

Especially after the year 2000, companies have made massive investments in re-engineering their business processes, and many of them realized that those investments need protection from events other than natural disasters, including outages, power failures, security incidents, and misconfigurations. New methodologies, techniques, and commercial products have been developed to manage backups and to speed up recovery. Thanks to these advancements, it is now possible to plan for a recovery time of just a few hours (still far, however, from the "zero downtime" dream sought by many and sold by others).

Nevertheless, many enterprises still have minimal contingency plans in place to ensure business continuity for their business requirements. Managers fail to plan for disaster recovery in the development life cycle of their projects, resulting in more risks and exposures than in the past. From a technological standpoint, some managers and planners still do not understand the intrinsic fragility of the new architectures for e-business to which they are migrating. As a consequence, they do not fully understand the increased requirements for business continuity, do no plan for the inevitable failures, do not mitigate the risks, and are not prepared to sustain the consequences.

In this chapter we address the problem of disaster recovery planning. The remainder of the chapter is organized as follows. First, we present the risks, that is, the main causes of failures that must be accounted for in disaster recovery planning. Then we discuss the possible costs that companies may sustain if they unsuccessfully respond to failures. We continue with illustrating the main phases of the disaster recovery planning process, including the risk management phase, to which a whole section is then devoted. The chapter then presents backups and alternate sites as main techniques for recovering data and providing for the continuity of the business processes in the occurrence of failures. We also discuss failure issues in the emerging Web-hosting service scenario. Finally, we provide a reference template for disaster recovery planning.

RISKS, OR THE CAUSES OF FAILURES

Failures in the IT infrastructure may have many causes that, with a broad classification, can be grouped as

Natural, such as, flood, fire, tornado;

Human, such as operator error, sabotage, network intrusion, malicious code, and denial of service; and

Environmental, such as equipment failure, software error, telecommunication network outage, and electric power failure.

Different failure causes differ in terms of *likelihood of occurrence* and *potential damage* that they can cause. For instance, events like sabotage, floods, or fire, usually considered in traditional disaster recovery, are rare but, when they happen, they turn out to be catastrophic. On the other hand, events like equipment failures or software errors are less catastrophic in nature, but may occur frequently and represent almost day-to-day problems for system administrators. Usually, these latter types of failures are not devastating for a company, and in past years (in the pre-Internet age), they were not included in disaster recovery planning. However, with today's companies relying more and more on networked systems, complex distributed multicomponent architectures, and the Internet as a primary business strategy, short but frequent downtime periods can become a major problem in terms of competition, reputation, and revenue losses (Kaye, 2001; Toigo, 2001b). For each possible cause of failure, likelihood of occurrence and possible damages must both be considered in disaster recovery planning.

Figure 1 shows the results of a first study mentioned in (Kaye, 2001) on the causes of IT failures.

According to this study, only 20% of downtime is caused by technology failures including hardware (servers and network devices), environmental factors (e.g., cooling and power outages), and natural disasters. By contrast, 40% of downtime is caused by application failures, which include bugs, operating system crashes, performance slowdown, incorrect updates, and changes to software. The remaining 40% of failures is due to human errors. In other words, according to these results, only 20% of downtime is caused by problems in the basic IT infrastructure, while the remaining 80% of failures is due to people and process issues, including software that provides for the business services and management operations (Kaye, 2001; Noakes-Fry & Diamond, 2001, 2002).

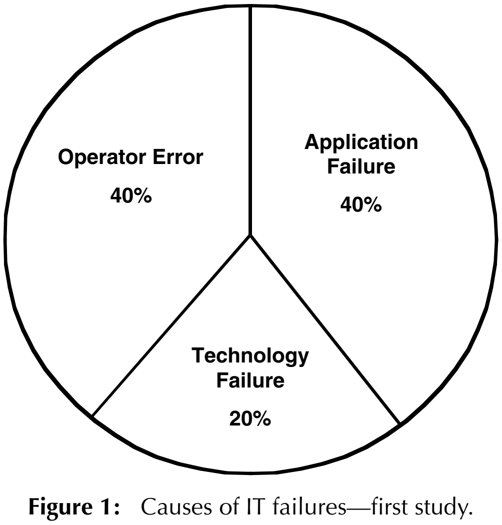


Figure 1: Causes of IT failures—first study.

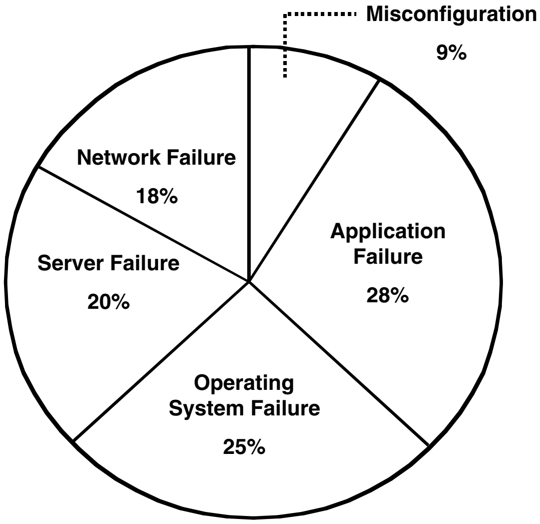


Figure 2: Causes of IT failures—second study.

A second study mentioned in Kaye (2001) investigates the technological components that fail. The results are illustrated in Figure 2.

According to this study, all the main components of the IT infrastructure, that is, network, servers, operating systems, and the applications, fail with comparable frequency, with a predominance of applications and operating systems. Yet, mere human errors or unskilled staff that lead to misconfigurations will cause a significant amount of outages (Kaye, 2001).

Although the two studies may appear contradictory, they are not, as they have adopted different classifications: the first study explicitly classifies human errors, while the second study includes them in the involved technological components. Merging the findings of the two studies, and reasoning on the different perspectives underlying them, can result in the statistics illustrated by Kaye (2001) as in Figure 3.

According to these consolidated results, the most frequent causes of outages are application bugs and connectivity failures. Applications frequently fail because of all sort of problems, including code bugs, bad design, weak requirement analysis, inconsistent tests, misconfigurations, erroneous installation or deployment, incompatibility with the run-time environment, and interference with other applications. As for connectivity failures, this last study estimates that misconfiguration of devices, routers, switches, or DNS should be expected to be higher than the 9% estimated in the study reported in Figure 2. Other frequent causes of outages are network partitioning and Internet service provider (ISP) failures. It is interesting to note how hardware failures represent only a small portion of the risks. As a matter of fact, there are now many reliable hardware vendors in the market and the

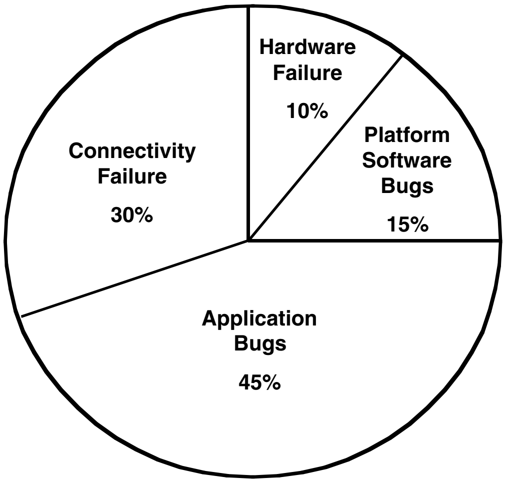


Figure 3: Causes of IT failures—consolidated results

quality of their products has improved in the years. The suggestion from experienced disaster recovery planners is then to select products and vendors with good reputation and just be skeptical about the latest news from the industry because they often provide exciting new features but at the price of less reliability. Similarly, also platform software (e.g., operating systems, DBMSs) is not one of the major risks, since products are generally reliable or could be configured and managed in order to maximize their reliability.

LOSSES, OR THE COSTS OF FAILURES

Internet-based companies, which require connectivity and availability of their Web sites 24 hours a day, 7 days a week, 365 days a year, have suffered all sorts of system and/or network outages in the past few years.

The cost of a disaster or a service interruption is not only the cost of the loss of business in the downtime frame. There are many associated factors that produce costs and losses, including

Brand image recovery;
Loss of share value;
Loss of interests on overnight balances;
Delay in customer accounting;
Loss of control over debtors;
Loss of credit control and increased bad debit;
Delayed benefits of profits from new projects or products;
Loss of revenue for service contracts;
Cost of replacement of building, plant, equipment, software;
Loss of customers;
Loss of profits;
Liability claims; and
Additional costs for advertising and marketing.

For instance, the brand value of most companies, while not reflected in their balance sheets, is one of their most valuable assets, reaching dozen of millions of dollars for some companies.

Also, statistics have demonstrated that the share price of a corporation suffering a disaster falls by around 5 to 8% within the first few days after a disaster. Recovery of the share price depends on the efficiency in recovering critical business functions. Past cases have shown that efficient recovery has let companies regain the confidence of financial analysts soon, and their share values have rolled back to the prices preceding the disaster, and even increased by 10 to 15% in the following 100 days. Companies that recovered with many difficulties, instead, were penalized with more share values falls, around 15% on average (Hiles, 2002).

A classification of losses should include (Kaye, 2001)

Fixed-asset losses. Fixed-assets of an IT infrastructure generally include the hardware and software used in the corporate business working. Fixed-asset losses can be caused by computer failures, fire, and thefts. Usually these risks can be mitigated through warranties, service contracts, and insurances.

Third-party liabilities. Depending on the type of business, a company can be exposed to third-party liabilities. These may be caused by an interruption in service provision or by the accidental disclosure of sensitive data after a security incident. Third-party damages are often covered by insurances.

Loss of intellectual property. The loss of data, application code, and sensitive information related to business processes can severely affect the company business. Backup and recovery strategies can be used to mitigate these losses.

Losses of revenues and profits. The loss of revenues and profits is the most evident type of risk in which a company may incur after a disaster or an outage. For companies that rely heavily on the Internet for their business, even short losses of connectivity or unavailability of their corporate portals can result in huge losses of revenue and profits.

As illustrated in Table 1, which shows some figures resulting from a study by META Group (DiNunno, 2000), depending on the type of industrial sector, the cost of downtime can reach noticeable values.

These figures show how important it is for companies to have a good disaster recovery plan in place. Statistics have shown that a typical organization may spend three or more times its normal annual marketing budget in the aftermath of a disaster to retain customer confidence and to retain or regain market share.

THE DISASTER RECOVERY PLANNING PROCESS

Disaster recovery planning consists in a set of activities aimed at reducing the *likelihood* and the *impact of disaster events* on critical business assets (NIST, 2002; Toigo, 1999; Wold, 2002). Disaster recovery planning is as complicated as any major system development project, with many stages of analysis and an overall activity encompassing

Table 1 The Cost of Downtime—(a) Loss of Revenue per Hour, (b) Loss of Hourly Revenue per Employee

(a)	**Revenue/hour**	**Industry sectors**
	$3,000,000 – $2,000,000	Energy, telecommunication
	$2,000,000 – $1,500,000	Manufacturing, financial institutions
	$1,500,000 – $1,000,000	Insurance, retail, pharmaceutical, banking, information technology
	$1,000,000 – $500,000	Food/beverage processing, consumer products, chemicals, transportation, utilities, health care, professional services
	$500,000 – $300,000	Electronics, construction and engineering, media, hospitality and travel

(b)	**Revenue/employee-hour**	**Industry sectors**
	>$1,000	Financial institution
	$1,000 – $500	Energy
	$500 – $200	Insurance, retail, utilities, construction and engineering
	$200 – $100	Telecommunication, food/beverage processing, consumer products, chemicals, transportation, health care, manufacturing, information technology, pharmaceutical, banking
	<$100	Professional services, hospitality and travel

all critical assets, business functions, and technological solutions.

Phases of Disaster Recovery Planning

A disaster recovery planning project begins by considering the *risks* and how to provide for their management so to make their occurrence tolerable to the organization (Noakes-Fry & Diamond, 2001, 2002; Toigo, 1999; Wold, 2002). The disaster recovery planning, which needs senior management approval to start, consists of (i) planning the actions to contain and recover from an emergency (e.g., procedures, teams, objectives) and (ii) preparing the agreements and the procedures for the collaboration among teams from different branches of the organization and with external partners. The collaboration procedures can be defined by a cross-functional team that includes top management, a purchasing representative, IT team members, maintenance personnel, and software user departments.

Another cross-functional issue that the disaster recovery planner should define together with top management and the public relation staff concerns possible interviews and news coverage. The importance of this issue depends on the exposure of the organization and can then vary depending on the business sector. In certain cases (e.g., financial institutions, government agencies or Internet-based companies) misleading press information can result in extremely high losses, affecting brand image and share values, and imposing costs for subsequent advertising/marketing campaign. Consequently, the disaster recovery plan should make sure that everyone in the organization understands who the media contact person is so that uninformed members of the organization do not make comments that are not in line with the public relations goals of the organization.

Once the project's development phase is completed, all procedures should be tested and personnel should be adequately trained. Testing helps identify possible weaknesses in the plan and evaluate the ability of recovery teams to implement the plan efficiently. Plan testing for the IT infrastructure should include (NIST, 2002).

System recovery on an alternate platform from backup media or mirrored sites;

Internal and external connectivity;

System performance using alternate equipment;

Restoration of normal operation; and

Notification procedures.

In addition to these IT-specific actions, all procedures for physical risk containment and recovery should be tested, including

Personnel evacuation plan; and

Alternate facility activation and recovery.

Training should make personnel aware of the disaster recovery procedure and provide those teams more involved in handling an emergency with sufficient experience to be able to execute their tasks with minimal support of documentation (as documentation might be unavailable in the first moment of an emergency). Plan elements that could be the subject of training include (NIST, 2002).

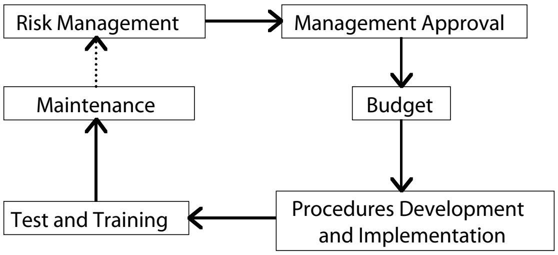


Figure 4: Disaster recovery planning project.

Purpose of the plan;

Cross-team coordination and communication;

Reporting procedures;

Security requirements; and

Team- and individual-specific processes.

When the plan is effective, it must be maintained, with periodic review, revision, and maintenance. Figure 4 summarizes the phases of a typical disaster recovery planning project (Kaye, 2001).

A disaster recovery planning should encompass all the IT infrastructure's components. In particular, all data repositories and all critical assets should be considered. More and more data are stored in a variety of supports and media, not just on mainframes and database servers; desktops and portable systems often hold important information (NIST, 2001; Takemura & Taylor, 1996). IT platforms to be considered in the planning should then include

Desktop computers and portable systems;

Servers;

Web sites;

Local area networks;

Wide area networks;

Distributed systems; and

Mainframe systems.

For each IT platform type, the *requirements* to recover each platform should be formally documented and considered in the analysis stage, and the *solutions* for addressing the event of a disaster should be detailed. The technical solutions should encompass *preventive* countermeasures (to avoid the failures) and *recovery* actions (to minimize the damage and re-establish the normal system operation in case of failures). Related documentation should describe

The frequency of backups and offsite storage of data;

The production and safe storage of an up-to-date copy of all applications and operating systems, together with their configurations, patches, and installation;

The redundancy of critical components;

The documentation of system installation, configuration, and requirements;

The interoperability between system components and between local and remote site equipment; and

The requirements for power management systems and environmental controls.

Business Impact Analysis

Convincing corporate management to invest in disaster recovery planning is often a greater challenge than dealing with the technical problems of backing up critical systems, minimizing the downtime, and maintaining the network connectivity. As it is often noticed, unfortunately, security investments usually receive low priority by managers, who prefer to invest in facilities and features that improve the business productivity, and realize the importance of protecting against disasters and a system's disruptions only after they happen. To complicate matters, the management decision of investing in disaster recovery capabilities is not a one time event, but must be sustained over the time, since every disaster recovery plan needs to be regularly tested, reviewed, and updated to be effective. Also, disaster recovery planning needs to be managed, improved, and updated with respect to the modification of the IT infrastructure and critical assets. For these reasons, the management consensus is vital for every disaster recovery plan that, in the best of circumstances, will provide for disaster avoidance by detecting and reacting to potential problems before they become disasters. Experienced disaster recovery planners usually observe that there are typical issues that should be justified to senior management and, among them, the most relevant is to carefully justify the plan's expenditures. Disaster recovery plans should be adequately documented with in-depth analysis of the risks addressed in the plan, the benefits for the corporation, the possible impact on business of a weak or insufficient disaster recovery plan, the estimates of the cost of downtime, and the possible reductions of insurance costs (Kaye, 2001; Toigo, 1999).

The impact of a disaster from a financial and an economic standpoint (Hiles, 2002) can be determined by means of a *business impact analysis* (BIA), whose objectives are to correlate each system component with the critical services it provides so as to characterize the consequences of the disruption of the component. A BIA therefore should

Identify critical IT resources. Different components, applications, services, and processes can have a different importance on the system services and functionalities. It is therefore important to determine priorities on the system functions and consequently on the resources for their support. Identification of critical resources includes PCs and portable systems (especially those of top management and system administrators).

Identify disruption impacts and sustainable time window in which recovery should take place. As we have already noted, timeliness is a key factor in recovery. The BIA must then identify, for each resource, the maximum time window for which the resource can be denied and the possible cascading effects that the outage can have on other processes. This includes identifying financial and nonfinancial costs associated with a disruption.

Develop recovery priority. Based on the criticality of resources and on the associated sustainable time window and cost consideration, recovery strategies can be prioritized so to optimize allocations and expenditures for the recovery process.

The BIA can provide input to the risk management process (see next section) on business risks that may not otherwise be identified. It can then be used to raise awareness of disaster recovery planning on senior management and focus individuals on their potential responsibilities, possible solutions, and costs.

THE RISK MANAGEMENT PROCESS

Risk is defined as the possibility of suffering harm or loss. Risk management is the process of planning for such eventuality. The goal of risk management is not to eliminate all risks (which would be impossible in practice), but rather to find an *acceptable balance between reducing risk and preparing the organization for losses*. The risk management process can be summarized in three steps:

Risk analysis: (i) identify the types of losses that may occur; (ii) determine the causes and the likelihood of possible losses; and (iii) estimate the resulting costs for the organization.

Risk mitigation: (i) identify actions to reduce or eliminate the likelihood of occurrence of each risk; (ii) determine the cost of those actions; and (iii) produce a cost/benefit analysis of the actions with respect to each risk.

Risk transfer: (i) Address the losses that may occur despite preventive actions and (ii) consider transferring them to insurers or third parties like service providers, vendors, or outsourcers in general.

We next examine each of these three steps in more details.

Risk Analysis

Risk analysis has two objectives:

Identify business processes and their associated IT infrastructure resources, such as the data, applications, systems, and networks used in supporting the corporate business. Business processes should be classified with respect to criticality in case of failure; and

Identify existing risks to the business processes and IT infrastructure.

Basically, risk analysis is concerned with identifying potential losses and the risks that might cause them. As already discussed, risks include failures of connectivity, hardware and software, as well as security incidents and natural disasters (Kaye, 2001; Toigo, 1999; Wold & Shriver, 2002). All the types of risks that can cause system disruption, namely natural, human, and environmental causes, should be analyzed and prioritized in disaster recovery planning. For each failure cause, a disaster recovery planner must define approaches to manage the risk and estimate the trade-off between sustaining the possible losses and implementing a disaster recovery strategy.

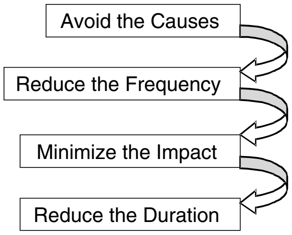


Figure 5: Risk mitigation process.

Risk Mitigation

Risk mitigation is a process aimed at limiting the likelihood of risks and the potential losses that risks can cause (Kaye, 2001). As sketched in Figure 5, the process is composed of the following phases:

Avoid the causes. Make conservative and prudential technology choices for critical assets: avoid technology risks by developing services in a robust manner and with the support of reliable platforms instead of adopting the latest technology (usually riskier), unless it has proven to give a sensible competitive advantage.

Reduce the frequency. All IT infrastructure components should be chosen among those ensuring a low rate of failure. Operating systems and software/hardware platforms of different vendors (or different versions produced by the same vendor) often have different failure frequencies. Reliability, and not just performance and features, should be carefully evaluated when technology is purchased to enforce the risk management strategy.

Minimize the impact. Since the frequency of failures and outages can never be reduced to 0, a risk mitigation strategy should analyze all possible single points of failure, whose failure may lead to major losses in terms of downtime and extent of service interruption. Redundancy is the typical strategy that can be applied to servers, network devices, and internal components.

Reduce the duration. Minimizing the duration of downtime is the ultimate goal of every disaster recovery plan: the more the downtime, the more the revenue loss. Many strategies can be adopted to speed up the recovery of data and functions. These may involve on-site hardware spares, on-site staff, efficient backup solutions, and alternative operative locations.

Risk Transfer

Risks that cannot be sustained by a company can be transferred to a third-party that assumes the risks in exchange of a service fee (Kaye, 2001). Hiles (2002) estimated that, on average, up to 40% of actual losses could be transferred, but no more than that. Risk transfer can assume two forms: *insurance* or *outsourcing*.

Insurance is the traditional way of transferring risks that cannot be mitigated with in-house solutions or transferred by purchasing the service outside. Most insurers offer coverage even for risks related with Internet-based threats like intrusions or denial of service attacks. With *outsourcing,* a third-party provides a specific service in place of an in-house solution. This third party can then assume the risk of losses that the company may incur by refunding the fee paid by the company (as in the case of Web-hosting services) or reimbursing the losses (as in the case of vendors). Risk transfer is established via a contract that usually includes a service level agreement (SLA) that should guarantee the company that purchases the service. This is a suitable solution in many situations, but the trade-off between the cost of possible downtime, if the service is realized in-house, and the cost of the purchased service with a certain quality level should be carefully evaluated. A careful evaluation of the return on the outsourcing investment should therefore be done in order to figure out whether the strategy is convenient (downtime costs versus service fees). Consider, for example, the case of Web-hosting services, where outsourcing is the common choice today. When a critical service is purchased from a vendor, instead of managing it in-house, it is typical to sign an agreement concerning the level of service guaranteed. In case of networked services, Web hosting, mainframe hosting, and so forth, the parties agree on a certain uptime guarantee, expressed as a percentage of time of availability (usually a value between 99 and 100%) of the purchased service. Clearly, the more the uptime guarantee, the more the cost of the service. However, how should a company decide which is the proper uptime guarantee value?

Kaye (2001) shows, as illustrated in Figure 6, the typical relationship between the amount paid for a certain service level and the value (i.e., the benefit that the company receives) of the increasing percentage of uptime guarantee (or, conversely, the decreasing downtime).

This reference relationship is meant to say that generally, after a certain percentage of uptime guarantee, the cost that a company would pay for a given service level is not worth the benefit gained. In fact, the revenue losses would be less than the cost paid and the investment would not yield a positive return.

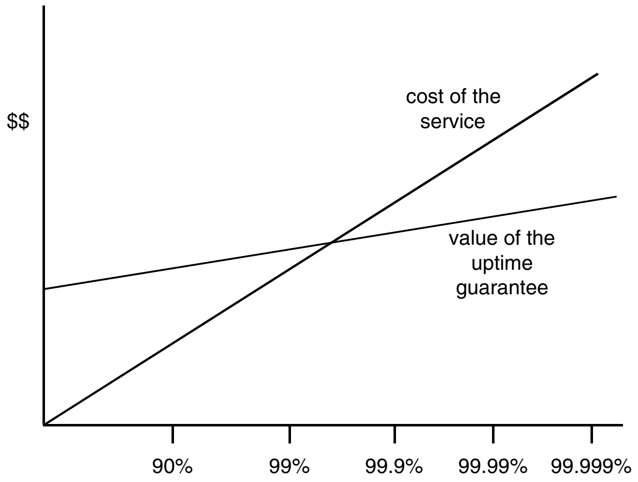


Figure 6: Uptime value vs cost.

Let us discuss this more with an example taken from Kaye (2001). Consider a company with $50 million in revenues and 15% pretax net profit. Increasing from 99 to 99.9% uptime will lead to increased revenues of $183,960 per month and profits of $27,594. If the cost of the service is less than $27,594, then the investment is worthwhile. If the company wants to upgrade from 99.9 to 99.99%, than the increase in monthly revenue is just $18,396 and $2,759 for profits. Hence, in this case the raise of the service cost should be much smaller, less than $2,759, to be convenient.

The example, clearly, is oversimplified since in practice a company should consider its business strategy and all the costs involved (e.g., loss of competitive advantage, market positioning, reputation, third-party liabilities) in order to evaluate the investment. However, the decreasing value of upgrading the service level is usually confirmed in most situations and should be carefully evaluated, possibly with formal models, to establish the return of the investment (Toigo, 2001b).

BACKUP AND RECOVERY

The goal of disaster recovery is to provide the ability of recovering the critical business functions of an organization in the occurrence of a disaster. One of the most important requirements for an effective recovery is the *availability* of *all the company's data* as well as of *the configuration of all critical components of the IT infrastructure* (e.g., backup of servers and network devices configuration, patches, passwords, routing tables, and firewall's rules). Although the need for maintaining backup copies of the data is well recognized, the need for a copy of all system configurations, in order to effectively restore the functionality of the IT infrastructure, is often overlooked. The reason why this requirement is so frequently overlooked is due to its intrinsic organizational nature: system management and day-to-day administration of the IT infrastructure (like changes in routing tables or firewall rules, or installation of patches) must be fully documented. Also, procedures should be put in place to make sure that every modification to system components (the critical ones, at least) is reported in a backup copy. Effective recovery requires also that *media for custom and purchased software* together with their *license information* should also be secured. Finally, the disaster recovery plan itself should be backed up. It may seem obvious that the disaster recovery plan will be available when an emergency calls for its activation, but this means that a complete copy of the disaster recovery plan should be kept in a safe place (e.g., offline at an alternate site). The need to back up all the information and elements needed to fully restore a system applies even when a comparable system is kept offline; the offline system maintains the exact level of operating system patches and operating system "state."

Some considerations should be devoted to backup requirements for desktop computers and portable systems, which have some peculiarities with respect to servers. In particular, when the PC data backup process is not automated (e.g., saving data on a fileserver to be regularly backed up), which is frequently the case, a user's duty of

Table 2 Example of Data Classification for Backup and Recovery

Classification	Definition
Critical	Data that must be preserved for legal reasons, for use in business-critical processes. Must be restored as soon as possible in the event of a disaster.
Vital	Data needed for important, but not critical, business processes. A loss of this data would represent severe damage and economic loss for the company. These data may have privacy requirements.
Sensitive	Data used in normal business operations, data that could be reconstructed from alternative sources in the event of loss but at some cost.
Noncritical	Data that could be reconstructed with minimal costs in the event of loss. No privacy or security requirements.

saving critical data on a backed-up server should be ruled by an appropriate usage policy (i.e., a document possibly subscribed by each user stating, among others, that each user should take care of saving his/her critical data on a backed-up server). The policy should be enforced by intermediate managers and users should be trained about the importance of this recovery requirement. Providing an automatic backup process and restoration for PCs requires *interoperability and platform standardization*. It is then usually recommended to have homogeneous platform operating systems, configurations, and applications among the organization's PCs and portable devices (NIST, 2002).

Backups should be kept in a safe place, where they cannot be affected by the consequences of a disaster, so to enable their *recovery* if the original is compromised or unavailable. For decades this has been the golden rule of disaster recovery, especially for mainframe-centric IT infrastructures. Traditionally, backup has been seen as a batch process running during night hours and storing on tapes all relevant information produced that workday. Although this solution is still used in a large number of organizations and is well suited in many circumstances, in many others cases this scenario is now anachronistic. With the increase of Web-based services and extranets, applications and business processes tend to be active on a 24 × 7 basis. As a consequence, there is no longer the notion of "night hours," which is meant to be that window of inactivity of business transactions perfectly suited to take a snapshot of data. In many real-world scenarios, transactions are always running and information is always accessed and possibly updated. In addition, data are growing every year at a terrific rate. Many commentators are commonly referring to the ever-growing amount of data to manage, back up, and restore as "the explosion of data," to stress both the impressive growth and the chaotic nature (Campbell, 2000; Kaye, 2001; Nicolett & Berg, 1999; Toigo, 2001a).

Backup and recovery planning should achieve two fundamental goals:

Backup should be made with a minimal impact on the production environment and, ideally, should be performed without blocking transactions execution.

Recovery should be efficient so as to minimize the time frame necessary to restore the data and bring the system back to its operational state. The efficiency of the recovery process is the key factor in the company's effective restoration of its business functions.

Backup and Recovery Techniques

The definition of data recovery strategies requires effective backup strategies to be in place. These, in turn, require the analysis of data to define priorities and criticality, so as to provide means to perform the backup process efficiently. The classification of business-critical data is needed to better organize their backup, optimize recovery, and therefore minimize downtime, all of which should follow the evolution of the company's business processes and the corresponding flows of information. Table 2 illustrates a simple typical data classification schema (Kaye, 2001).

To better maintain a clear identification and classification of data, a formal policy should be adopted. The goal of the policy is to establish who is in charge of formally classifying data for each business unit, department, or workgroup; and define a uniform classification methodology for the whole organization corresponding to the recovery priorities and requirements in the event of a disaster.

Some of the main strategies and technologies for backing up data are summarized by Toigo (1999) as follows:

Server image backup. Image backup creates a physical image of an entire disk. It operates at physical disk level by transferring sectors of physical data storage blocks from hard disks to tapes. It has fast transfer rates and provides a complete server backup by transferring the entire content of a hard disk as a single unit. This technology requires the system to be inactive during the backup.

Snapshot or versioning backup. This technique is fast because after an initial complete image backup, it proceeds incrementally by backing up the modifications only. It works with block-level copies on a partition basis. It is widely used in traditional tape-based backup systems.

Full-volume backup on a file-by-file basis. Complete data sets are copied over the production network to backup devices. A process controlled by a server initiates the transfer to backup devices.

Incremental or differential file backup. Files are compared with their version in the preceding backup and only changed files are backed up again. *Differential* backup is meant to keep separate versions of the same backed up files, while *incremental* backup just keeps the last versions.

Backup using object replication. Logical objects are defined by software, including disk partition tables, boot volume, security information, system volume, and user data volumes. Logical objects are regarded as a single logical unit and all data can be copied at once.

Electronic tape vaulting. It replaces the manual procedures for handling tapes stored in offsite storage facilities and for moving tapes between locations. It exploits a wide area network connection between the company facility and the offsite storage facility where tapes are stored to provide for immediate access to data stored on tapes. Technologies like shadowing and multiplexing controllers allow duplicating backup data streams in two or more tape devices simultaneously. This way a company could make both a local backup (for complete recovery in the event of a disaster) and an electronic vaulting on a remote storage facility (to preserve data from disaster, like flood or fire, affecting the company site).

Remote disk mirroring. It is another alternative to physically handling tapes. With remote mirroring, the content of disks is duplicated in near real time to a remote disk volume. The same concept is applied to RAID (redundant array of inexpensive disks) technology as a means of fault tolerance instead of disaster recovery.

Selecting an *offsite storage* facility is the safest choice for preserving backups. The remote storage facility should be chosen because it will be unaffected by the same disasters that might affect the main company site. The remote site could be a branch of the company or a commercial system recovery facility offering this service. While companies with decentralized sites could choose to keep backups among their different sites, several commercial storage facilities offer backup storage as a service, including media transportation, safe storage, and recovery (NIST, 2002; Nicolett & Berg, 1999; Toigo, 1999). When selecting an off-site storage facility and vendor, there are some key factors to keep in mind:

Geographic area. Distance from the organization and the possibility of the storage site being affected by the same disaster as the organization (locations geographically close to the organization are likely to be affected by the same natural disasters).

Accessibility. Length of time to retrieve the data from storage and the storage facility's operating hours.

Security. Security capabilities of the storage facility and employee confidentiality, accordingly to the data's sensitivity and security requirements.

Environment. Structural and environmental conditions of the storage facility, like temperature, humidity, fire prevention, and power management control.

Cost. Cost of shipping, operational fees, and recovery services.

ENSURING CONTINUITY OF OPERATIONS: ALTERNATE SITES STRATEGY

Disaster recovery plans have traditionally accounted for natural disasters as the first threat to an organization. Hence, beside the safe storage of data backups, it has been necessary to provide an organization with a plan for the continuation of its business functions while severe damages occurred to its site and hardware assets. A disaster recovery plan should then include a strategy to recover and perform system operations at an alternate facility. Generally, three types of alternate site strategies are possible (NIST, 2002):

A dedicated site owned or operated by the organization;

A reciprocal agreement with an external entity (e.g., a business partner); and

A commercially leased facility.

As prospective alternate sites are evaluated, the disaster recovery planner should ensure that the system's security, management, operational, and technical controls are compatible with the prospective site. Controls may include measures for network security, such as firewalls and intrusion detection systems, and physical access controls. If the service is purchased from a vendor, an SLA must be negotiated carefully. The SLA must clearly state all conditions like testing time, workspace, security requirements, hardware requirements, telecommunication requirements, support services, and recovery time frame. The disaster recovery planners should be aware that a commercial facility usually hosts many customers, which may all be simultaneously affected by the same natural disaster. Hence, the commercial facility could be asked to recover different customers simultaneously and it may not have enough resources to do it properly. In this case, the commercial facility would tend to set priorities among its customers and not treat all of them at the same level of service. Unfortunately, this scenario is realistic and for this reason it is extremely important to negotiate a clear and extremely detailed SLA with the commercial vendor, in order to set once and for all how possible disasters should be addressed and how priority status is determined. Regardless of the type of alternate site strategy, the alternate facility could provide different services and have a different level of readiness with respect to the goal of immediate resumption of all business functions. There is a typical classification that also reflects the different commercial services offered on the market (NIST, 2002; Nicolett & Berg, 1999; Toigo, 1999):

Cold site. It typically consists of a facility with adequate space and infrastructure (electric power, telecommunication connections, and environmental control) to support the IT infrastructure. The cold site contains neither IT equipment nor office automation equipment, like telephones, fax machines, or copiers. The organization leasing a cold site should provide for all the necessary equipment.

Table 3 Alternate Site Characteristics

Site	Cost	Hardware equipment	Telecom	Setup time	Location
Cold	Low	None	None	Long	Fixed
Warm	Medium	Partial	Partial/full	Medium	Fixed
Hot	Medium/high	Full	Full	Short	Fixed
Mobile	High	Dependent	Dependent	Dependent	Not fixed
Mirrored	High	Full	Full	None	Fixed

Warm site. It is a partially equipped office space containing some hardware, software, telecommunication, and power sources. The warm site is kept in an operational status to be promptly made fully operational in the event of a disaster at the company's main site.

Hot site. It is a fully functional and immediately ready facility equipped with all the needed hardware, software, telecommunication, and power sources. Hot sites are usually staffed 24 hours a day, 7 days a week. Recovery of business functions is guaranteed to be extremely fast.

Mobile site. It is a self-contained, transportable shell equipped to restore immediately the subset of the most critical business functions, including telecommunication facilities. Major commercial vendors lease these mobile sites and keep them located in strategic places to let them reach most of the industrial plants in a relatively short time. However, to guarantee an on-time recovery by means of a mobile site, a company should sign agreements with vendors to have the mobile site ready in a time frame compatible with the business requirements.

Mirrored site. It maintains an almost exact replica of the IT infrastructure of the company's main site and is always ready to recover all the business functions. Recovery is almost immediate through a mirrored site. It is often owned and operated by the company, given the high degree of customization and specificity.

Table 3 summarizes the characteristics of the different alternate sites as classified in NIST (2002).

The solutions listed in Table 3 differ in terms of cost and timeliness. Cold sites are less expensive to maintain but require considerable time to acquire and install the necessary equipment. Going down the list, we find solutions that result in more expensive costs but provide greater availability. Mirrored sites are the most expensive solution and provide almost 100% availability (i.e., minimal downtime). Remote mirroring has gained large popularity as a reliable solution for disaster recovery; we therefore discuss it in more detail next.

Remote Mirroring

Advances in networking technology have put the basis for many commercial systems on remote mirroring for safeguarding corporate data and providing fast recovery. However, sometimes the expectations that planners place on remote mirroring solutions are not realistic. Also, solutions based on remote mirroring tend to be complex, both architecturally and in data management, to satisfy to the high requirements from companies. This may lead to a lack of robustness of the recovery system, misconfiguration, and improper array maintenance. The consequence could be the loss of synchrony between arrays or the storage of corrupted data. If this happens, in the event of a recovery, the result could be a multiplication of the disaster, since the risk is to recover corrupted or incoherent data (Kaye, 2001; Toigo, 2001a, 2001b). Also, remote mirroring does not provide for zero downtime, which is considered a requirement by some senior managers. Remote mirroring, if configured properly and implemented with high-level technology, could be very fast, much faster than many other recovery alternatives, but it cannot reduce the downtime to 0. Hence, remote mirroring is not the panacea that makes the nightmare of downtime disappear: risk management and disaster recovery planning are still mandatory. Moreover, remote mirroring is subject to data gaps, since some sort of asynchrony is usually implemented (e.g., the mirroring between the secondary and the tertiary arrays of Figure 7). Possible consequences of incoherent recovered data should be carefully analyzed in the planning stage. These may provoke the loss of commercial orders, payments, or shipments, if data gaps are not identified and handled properly. Solutions designed to be completely synchronous exist but come at a high price in terms of performance and system latency. In fact, the problem in remote mirroring has been traditionally the latency caused by the distance, possibly many hundreds of miles, between the sites, in particular when the mirroring is synchronous (i.e., writes to production disks waits for the mirrored writes to be completed and acknowledged).

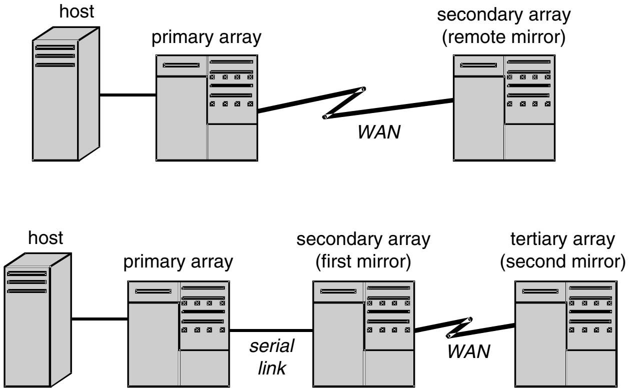


Figure 7: Remote mirroring configurations.

Starting from the mid-1990s, the cost of high-speed bandwidth WAN decreased and vendors realized those architectures that made remote mirroring one of the best solution at affordable costs for many companies. The typical improvement was to provide for three arrays, adding an intermediate site between the company and the final remote storage facility (Toigo, 1999). Figure 7 shows the traditional and the enhanced schema for remote mirroring. The benefit of the enhanced schema is twofold. First, the link between the company site and the secondary array could be a high-speed serial one, thus reducing the latency in synchronous mirroring. Second, the mirroring between the secondary and the final arrays, connected with a WAN, could be realized asynchronously, hence optimizing the mirror operation, while not affecting the robustness of the disaster recovery solution.

BACKUP AND RECOVERY FOR WEB-BASED HOSTING SERVICES

Web-hosting service (WHS) providers are commercial vendors that may store hundreds of Web sites in their systems. Moreover, the service effectively hosted may be complex, not just a Web site. WHS can be composed of *n*-tier architectures with Web portals, application servers, middleware layers, and community services. Due to this increasing complexity, nowadays it has become common for an organization to purchase the Web management and hosting service instead of implementing it in-house. However, like other outsourced services, even for Web hosting, customers should evaluate carefully how the service is provided and consider which guarantees the vendor provides in the event of an outage (Scott, D., & Natis, Y., 2000).

The primary deployment options for Web-based applications are *Internet data centers*. A growing number of enterprises are considering dual-site application architectures, which split application traffic between two active sites. The active sites could be corporate data centers, commercial data centers, or both. One aspect to consider carefully when services from commercial data centers are purchased is the large number of simultaneous recoveries they must provide for all their customers in the event of an outage. Thus, similarly to what we observed for remote mirroring, the customers of a Web-hosting service should make sure that the outsourcer has a robust disaster recovery plan and, through clear service level agreements, that she has a contractual responsibility to effectively execute it (Kaye, 2001; Toigo, 2001b).

Different technological options can be used for storage management of Web sites and Internet applications (Kaye, 2001):

Direct attached storage (DAS). The disk drives or arrays are directly connected to an individual server.

Capacity: ≤ 200 GB

Features: Well-suited for few Web servers with infrequent updates. Good reliability using RAID for databases.

Network attached storage (NAS). High-performance general-purpose file servers connected to Web sites via LANs.

Capacity: ≤ 10 TB

Features: Well-suited for several Web servers with frequent updates. Very good reliability for databases, if clustered.

Storage area networks (SAN). High-performance special-purpose storage systems connected to Web servers via dedicated fiber optical links.

Capacity: 100 GB–100 TB

Features: Not economically convenient for Web servers. High reliability for databases.

Global storage systems (GSS). A technology for geographically replicated file systems.

Capacity: 300 GB–100 TB

Features: Very good for Web accesses and frequent updates. Not well-suited for databases.

Networked storage solutions, however, could pose special difficulties that might significantly degrade their speed to that of some tape-based solutions. For instance, in a SAN, domain servers, routers, and software delivering virtual volumes to SAN servers form a networked infrastructure that poses an increased management burden on the IT staff and is more prone to failures than simpler solutions.

Most e-business applications hosted in Internet data centers are based on *n*-tiers architectures. These multitier platforms typically include one or more Web servers, application servers used to integrate Web technology with legacy systems, connectors and adapters, database servers storing data, and authentication and authorization servers filtering the user accesses and incoming connections. This complex architectural design and the interconnection of several components are clearly more prone to failures than centralized monolithic systems. These failures, often implying short but frequent outages, must be handled carefully in the planning stage or in negotiating SLAs with vendors. Selection of components and replication strategies should be required with the goal of minimizing downtime in mind.

In addition to this class of system failures, e-business needs to face the growing rate of network attacks and denial of service. Many commentators are now commonly reporting "the convergence of disaster recovery and network security" for e-business applications, meaning that the two areas have much in common and more and more interdependencies. From a disaster recovery planning viewpoint, intrusions and security incidents are another cause of downtime to be considered in the risk management stage, in the BIA, and in the selection of vendors. Internet data centers should provide for a staff skilled in network security and for robust security architectures and policies (Kaye, 2001; Toigo, 2001a, 2001b).

DISASTER RECOVERY PLANNING TEMPLATE

This section provides a reference template for disaster recovery planning derived from Brooks, Bedernjak, Juran, and Merryman (2002). It is only intended as one possible example of a disaster recovery planning template, since an actual plan should meet the peculiar requirements of each organization. Further examples and interesting readings for the definition of a disaster recovery plan could be found at Toigo (1999) and DRJ (2002).

A plan could have the following schema:

Introduction

Purpose. The purpose of the plan should be shared and fully understood by the personnel involved. For example, recall that in the event of a disaster, the plan provides for precise procedures to activate the plan, to recover from the disaster, and to restore normal operations. Other key actions to be introduced are the identification and classification of assets and services, the assignment of responsibilities and roles to personnel to be carried out during the interruption of normal operations, and the coordination required with other staff involved in the recovery phase, e.g., vendors and local or government agencies.

Scope. The scope of the plan describes which measures the organization has put in place in the event of a disaster. For example, which alternate site strategy has been set, whether a hot site is ready to recover critical operations, how long the alternate facility could be effectively used to back up the IT infrastructure, and how long could be the estimated unavailability of the main IT functions until activating the alternate facility (this is strictly dependent on the business sector so the range could stem from 48 hr or more of inoperability to immediate switching to a mirrored site in case of a failure).

Assumptions. The introduction could even highlight what will need to be fully operational and organized for the plan to be successful. For instance, it should be assumed that preventive safety measures must be followed, in particular in the case of a natural disaster. Thus, the sprinkler system, fire extinguishers, generators, etc. must be fully operational and properly maintained. Assumptions related to the IT infrastructure usually refer to the availability of the UPS (uninterruptible power supply) to back up a temporary power failure and to the safety of data backups in a site not affected by the disaster. Other assumptions could be related to the effectiveness of SLAs with service providers, e.g., which actions are to be carried out by the providers, which resources should be available, and how communication should be done. Of overall importance is the availability of a consistent copy of the configuration of all critical components of the IT infrastructure, which must be backed up in a safe place.

IT Infrastructure and Responsibilities

System Description and Architecture. A clear and comprehensive documentation about all the critical components and functions of the IT infrastructure is necessary to speed up the recovery after a disaster. The IT architecture should be described with details about the physical location of critical servers and network devices, the functionality of each component, LAN/WAN/SAN topologies, and telecommunication connections. Also, for each critical component all the information regarding their operating systems and applications should be immediately available. The plan should indicate the location where the backup copies, with software and actual configurations, are safely stored.

Responsibilities. Key roles in the organization's chart in the event of a disaster should be indicated, as long as the line of succession (e.g., if the Chief Information Officer is unable to function, his/her functions must be delegated to someone else). Aside from this, the plan describes roles and responsibilities of all the teams specifically dedicated to contain the effects of a disaster, organize the recovery of critical functions, and restore the normal operations. For each team, its members, leadership, roles, responsibilities, and coordination with other teams should be indicated.

Activation

Damage Assessment. The first duty of a disaster recovery plan is to contain the damages in the initial phase of an emergency. The top priority, especially in the event of a natural disaster, is the safety of the personnel; hence proper evacuation plans and directives must be enforced. Then the plan should state detailed procedures to determine the cause of the disaster, the extent of damages, the affected physical area, and the status of physical infrastructures and of the IT equipment. After a first inspection, a disaster recovery team should proceed to a detailed inventory, following the priorities set for the organization's assets.

Criteria for the Activation. Based on the results of the damage assessment phase, the disaster recovery coordinator could decide immediate actions like notify civil emergency personnel, the provider of the alternate facility, or the vendors for replacement items. The plan should define criteria for the activation. For example, how long the facility is estimated to be unavailable and how long until critical components of the IT infrastructure can be replaced. If the plan is to be activated, the coordinator should inform all team leaders, which in turn should activate all team members.

Recovery

Priorities. The key factor in the recovery phase is to follow priorities. Unless the disaster is confined to a single or few components, the recovery should be focused on the most critical assets and business functions. The goal of restoring as fast as possible those functions necessary for the organization business should be clearly stated and supported. Priorities have been already set during the risk analyses and the business impact analyses. The results of these two key phases of the disaster recovery planning should be reflected in primary, secondary, and tertiary recovery objectives that the disaster recovery teams should pursue.

Return to Normal Operation

Original or Alternate Site Restoration. This phase of the plan presents the activities required for restoring operations at the original or at the alternate site. Technical details may include which teams operate first and on which components, coordination activities among teams (for example, between network and system administrators, between system administrators and application administrators, and/or between personnel of the organization and personnel of the provider). Testing procedures could be also described.

Deactivation. Procedures should outline how to return to normal activity, like how to gracefully deactivate the alternate site or temporary measures put in place, and how to restore a proper backup system and the safe storage of backup data.

Appendices

Documentation. The appendices of the plan should include all the detailed documentation mentioned in the plan itself and needed during an emergency, for example, personnel contact list, vendor contact list, equipment and specifications, SLAs, IT standard operating procedures, BIA, evacuation plan, equipment inventory, and configuration data.

CONCLUSION

The scope of disaster recovery has become wider in recent years due to a change in the business functions of many enterprises. Internet connections have driven the development of new business models that require immediate recovery and continuous operations and have "zero-downtime" as their ultimate goal. While the "zero-downtime" goal is still impossible to achieve, and this should be stated clearly in each disaster recovery project, the disaster recovery area has developed new planning methodologies and techniques to match these requirements. As a consequence, the process of putting into place an effective disaster recovery plan has now sophisticated technologies to exploit but it is also more complex for its interconnections with both the IT infrastructure and the business functions of a company. An ineffective disaster recovery plan could result in severe losses when one of the many risks becomes reality. Such risks today are not only the traditional disasters like fire, floods, or sabotage, but also the "normal" failures of the IT infrastructure, such as misconfigurations, failures of applications, and security incidents, which may lead to frequent downtimes. E-business models can no longer sustain those frequent downtimes or network outages as in the past when companies could accept recovery-time of some days in the event of an IT failure.

Processes that allow achieving a better efficiency of the recovery phase in the event of a disaster, such as the risk analysis, the business impact analysis, and the innovative backup and recovery techniques, assume therefore great importance.

GLOSSARY

Alternate site A location, other than the normal facility, used to process data and conduct critical business functions in the event of a disaster.

Business continuity planning A generic term covering both disaster recovery and business resumption planning.

Critical functions Business activities that, if unavailable, can result in high costs for an organization.

Disaster Any event that can cause the inability of an organization to provide critical business functions for a certain period of time.

Disaster recovery planning The preparation of the activities and measures necessary to minimize losses and ensure the continuity of critical business functions in the event of a disaster, synonymous to contingency planning, business resumption planning, corporate contingency planning, business interruption planning, and disaster preparedness.

Disaster recovery teams A structured group of teams responsible of controlling the recovery operations in the event of a disaster.

Electronic vaulting The transfer of data to an offsite storage facility via a communication link or a different portable medium.

Facility A location containing equipment and infrastructure to perform normal business functions.

Loss The cost resulting from a disaster.

Network outage The interruption of system availability as a result of a communication failure affecting a network.

Offsite storage facility A secure and remote location where backups of hardware, software, data files, documents, or equipment are stored.

System outage An unplanned interruption in system availability resulting from operational problems or hardware/software failures.

CROSS REFERENCES

See *Computer Security Incident Response Teams (CSIRTs); Guidelines for a Comprehensive Security System; Physical Security.*

REFERENCES

Brooks, C., Bedernjak, M., Juran, I., & Merryman, J. (2002). DR and business impact analysis planning templates. In *Disaster recovery strategies with Tivoli storage management* (Appendix A) (IBM Redbook SG24-6844-00). Retrieved October 10, 2002, from http://www.redbooks.ibm.com/redbooks/SG246844.html

Campbell, R. (2000). *Continuity planning in the new millenium—The convergence of disciplines*. Retrieved October 10, 2002, from http://www.disaster-resource.com/cgi-bin/article_search.cgi?id='16'

Di Nunno, D. (2000, November). *IT performance engineering & measurement strategies: Quantifying performance loss* (META Group Research Delta Summary 142). Retrieved October 10, 2002, from http://www.metagroup.com/cgi-bin/inetcgi/jsp/displayArticle.do?oid=18750

DRJ (2002). DRJ's sample DR plans and outlined. *Disaster Recovery Journal*. Retrieved October 10, 2002, from http://www.drj.com/new2dr/samples.htm

Hiles, A. (2002). *Business impact analysis: What's your downside?* Retrieved October 10, 2002, from http://www.rothstein.com/articles/busimpact.html

Kaye, D. (2001). *Strategies for Web hosting and managed services*. New York: Wiley.

NIST (National Institute for Standard and Technology). (2002, June). *Contingency planning guide for information technology systems* (NIST Special Publication

800-34). Retrieved October 10, 2002, from http://csrc.nist.gov/publications/nistpubs/800–34/sp800–34.pdf

Nicolett, M., & Berg, T. (1999). *Storage backup and recovery for distributed systems* (Strategic Analyses Report R-09-3148). Stamford, CT: Gartner Group.

Noakes-Fry, K., & Diamond, T. (2001). *Business continuity recovery planning and management: Perspective* (Technology Overview DPRO-100862). Stamford, CT: Gartner Group.

Noakes-Fry, K., & Diamond, T. (2002). *Business continuity planning software: Perspective* (Technology Overview DPRO-100469). Stamford, CT: Gartner Group.

Scott, D., & Natis, Y. (2000). *Building continuous availability into e-applications* (Research Note COM-12-1325). Stamford, CT: Gartner Group.

Takemura, R., & Taylor, R. M. (1996). *The increasing need for client/server contingency planning*. Retrieved October 10, 2002, from http://www.disaster-resource.com/cgi-bin/article_search.cgi?id='36'

Toigo, J. W. (1999). *Disaster recovery planning: Strategies for protecting critical information assets* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall PTR.

Toigo, J. W. (2001a). Fighting fires on the Web. *Enterprise Systems*. Retrieved October 10, 2002, from http://esj.com/features/article.asp?EditorialsID=60

Toigo, J. W. (2001b). Storage disaster: Will you recover? *Network Computing*. Retrieved October 10, 2002, from http://www.networkcomputing.com/1205/1205f1.html

Wold. G. H. (2002). Disaster recovery planning process—Part I, II, and III. *Disaster Recovery Journal*. Retrieved October 10, 2002, from http://www.drj.com/new2dr/w2_002.htm, http://www.drj.com/new2dr/w2_003.htm, http://www.drj.com/new2dr/w2_004.htm

Wold, G. H., & Shriver, R. F. (2002). Risk analysis techniques—The risk analysis process provides the foundation for the entire recovery planning effort. *Disaster Recovery Journal*. Retrieved October 10, 2002, from http://www.drj.com/new2dr/w3_030.htm

Distance Learning (Virtual Learning)

Chris Dede, *Harvard University*
Tara Brown-L'Bahy, *Harvard University*
Diane Ketelhut, *Harvard University*
Pamela Whitehouse, *Harvard University*

INTRODUCTION: A CONCEPTUAL FRAMEWORK FOR DISTANCE EDUCATION

Understanding the likely trajectory of contemporary distance education requires envisioning today's virtual learning applications in a larger context. Organizational structures and processes for learning across distance are shaped by evolving societal needs; these include developing human skills for economic prosperity and political participation, increasing educational opportunity for disenfranchised learners, and preparing students for a technology-intensive workplace. Demographic changes and shifting student characteristics also are influential in forming the nature of distance education. The "typical" student using technology to surmount geographic barriers has shifted from a rural resident in an agricultural society seeking to understand industrialization to a huge proportion of the population using formal and informal means of education to participate in an emerging knowledge-based, global civilization. Technological evolution also molds distance education, which has incorporated successive generations of media that include lantern slides, radio, television, computers, telecommunications, and now cyberspace and ubiquitous computing. The Internet itself can be viewed as an extension of the technologies that sparked early forms of distance education over 150 years ago. Envisioning the future of virtual learning requires synthesizing all these factors, organizational, demographic, and technological, into ways sophisticated computers and telecommunications can further enable reinventing teaching, learning, and schooling.

HISTORICAL FOUNDATIONS OF MODERN DISTANCE EDUCATION

Pre-1900 Distance Education

In the 19th century, distance education in the United States was shaped by new technologies that allowed educators to overcome barriers of distance and time. These advances also altered beliefs about the purpose of education through changes in the social, political, and geographic dimensions of learning. The Progressive Era brought a shift in educational thinking from the Jeffersonian ideal of forming good citizens (Tyack, 1967) to Horace Mann's assertion that tax-supported education would provide literacy skills to enhance the employment opportunities of workers, as well as to "prepare citizens to fulfill their civic duties" (Cuban, 2001, p. 8). The roots of pre- and early 1900 distance education lie in this Progressive notion that children need education to become good citizens and contributing workers, while adults need ongoing education to improve their economic lives.

Pre-1900 forms of distance education were influenced by the need to reach learners in remote geographic locations, rural free delivery, and the birth of the Chautauqua Movement in the 1840s. This movement promoted continuing education for adults based on the ideal of lifelong learning and offered correspondence courses combined with some face-to-face meetings for adults to earn certification in certain subjects (Scott, 1999). The development and implementation of the first correspondence courses were credited to Sir Isaac Pitman of England, the inventor of shorthand. In 1840, he used the postal service in England to reach learners at a distance. Around the same time, in the United States traveling lyceum lecturers crossed the country with lantern slides, making possible national conversations between intellectuals and the citizenry. Ralph Waldo Emerson was a very popular lyceum speaker, and his talks generated many debates about the meaning of democracy (Field, 2001). A more formal version of the early American correspondence course was created by Anna Ticknor of Boston in 1873. To increase educational opportunities for women, she originated the Society to Encourage Studies at Home. The society provided courses of study for women of all social classes and served over 10,000 women over its 24-year lifespan (Nasseh, 1997; Stevens-Long & Crowell, 2002).

In 1878, John H. Vincent, co-founder of the Chautauqua Movement, created the Chautauqua Literary and Scientific Circle. This Circle offered a four-year correspondence course of readings; students who successfully completed the course were awarded a diploma. This course was open to all adults, including women and senior citizens (Scott, 1999). By 1892, the 19th-century version of the "information superhighway" (otherwise known as rural free delivery) paved the way for Pennsylvania State University to provide higher education to rural families (Banas & Emory, 1998, p. 365).

Other institutions of higher education, notably the University of Chicago and the University of Wisconsin, modeled their extension schools after the Chautauqua program (Scott, 1999).

In the pre- and early 1900s, distance education provided access to educational services for nontraditional learners and for those whose geographic or social location made continuing their education difficult. Students in remote rural areas, women, senior citizens, and others benefited from the Progressive ideals of access to education for all, as well as from the concept of lifelong learning, the founding notion of the Chautauqua movement. New technologies and services (e.g., lantern slides, railroads, rural free delivery from the postal service) provided the infrastructure to support the Progressive educational ideals of the times.

1900–1960 Distance Education

While distance education is rooted in the 19th century, the field blossomed in the 20th century. Distance educators look to technological innovations to provide new opportunities for their field, and the 20th century was rife with technological advances (Mood, 1995). During the 20th century, distance education embraced radio, television, computers, and ultimately the Internet. As the methods of delivery for distance education expanded, so did the diversity of learners seeking distance education and their reasons for enrolling in such courses. Individuals interested in learning cultural norms, becoming more capable in the workforce, or hoping to resituate themselves in their social context after wartime service became major consumers of distance education (Dymock, 1995).

As the field grew, governments discovered the potential of distance learning for spreading their messages and provided increased financial support (Sumner, 2000; Tait, 1994). Starting with the creation of the First International Council for Distance Education in the 1930s (Bunker, 1998), the formalization of distance education as a field of study established it as a valid method of education. However, only in the latter half of the 20th century did distance education gain widespread acceptance within the educational community, the result of a substantial body of research that confirmed its benefits (Nasseh, 1997).

At the start of the 20th century, distance education still relied on correspondence courses delivered primarily through the postal service. Many of these courses were delivered to their students by mail, as discussed earlier, but did not allow much interaction or individualization (Moran, 1993). Although rules for home study were established in 1926 to allow some form of governmental control, correspondence methods were not conducive to supporting learners nor were they standardized (PBS, 2002). Even the addition of motion pictures in the early years of the century failed to revolutionize this traditional pedagogy for distance education (Jeffries, 2002; Nasseh, 1997). One of the main goals of the early distance education programs was to help inculcate immigrants into the "American way of life" (Sumner, 2000), but these learners needed substantial guidance and aid. As a result, poor curricular design and lack of support were particularly problematic, and the dropout rate was high (Shea & Boser, 2001).

In the 1920s, distance education started to utilize radio for delivery of lessons (Bourke Distance Education Centre, 2002; Nasseh, 1997). In a push to widen access, speed the interaction between student and professor, and personalize the delivery of distance education, the use of radio was seen as an exciting opportunity. In the mid-1930s, an American art history course was offered by radio broadcasts (Funk, 1998), and other courses supported forming "listening groups" to enhance learning (Mood, 1995).

However, despite the rapid rise of radio technology, distance education courses were rarely if ever offered for credit in higher education (Nasseh, 1997; PBS, 2002). The education community, along with society as a whole, regarded legitimate education as only possible in conventional locales, such as classrooms (Funk, 1998). This view was clearly articulated by Harper (a designer for the Chautauqua program) in the late 1800s (Froke, 1995). He viewed the lack of face-to-face instruction to be an insurmountable problem in distance education, a view held even today.

To address this lack of teacher interaction, a modification of the correspondence course was designed in Soviet Russia in the 1930s, called the "consultation model" (Tait, 1994). As its name implies, this type of correspondence course, while primarily at distance, included periodic face-to-face meetings with instructors. However, unlike its name, the "consultations" were mostly lecture-based meetings intended to spread communist dogma. This model did offer an opportunity to put a face with the words, an issue with which today's distance education continues to grapple (Roberts, 1996).

Television was the next big advance in distance education technology. As early as 1934, the State University of Iowa used television to deliver course content (PBS, 2002). Early research into learning via television indicated mixed results, with several studies showing that it was similar to conventional instruction. Gayle Childs referred to televised distance education as an "'instrument' of delivery, not a pedagogical method" (Jeffries, 2002). Echoes of current debates about technology as a tool for learning were heard regarding the use of broadcast media in distance education (Nasseh, 1997).

Thus, the early 1900s were characterized by changes in both how distance education programs were delivered and why. New technologies such as radio and television increased the speed of delivery and personalization of distance education, and the students broadened to include those wishing to become steeped in a new culture. However, distance education was still plagued with problems and lack of acceptance by the established education community.

1960–Present Distance Education: The Rise of Computer-Based Learning

Prior to the introduction of computer technologies in the 1960s, correspondence course and independent study models of distance education posed challenges to the learning and teaching processes. This contributed to a persistent problem of credibility for the field. "Telecourses" (Verduin & Clark, 1991), developed in the 1970s, showed promise for minimizing some of these problems. Previously, television had primarily been used as an electronic blackboard and for the delivery of standardized content, through lectures intended to reach wide audiences. The development of videotape allowed educators to customize the same content for different learning environments. This medium also allowed increased flexibility; course content could be stored, delivered, and repeated at will. This minimized time-dependency, a drawback of previous televised courses. However, despite their advantages, the cost and complexity of producing telecourses made them impractical for teaching large numbers of students.

Around the same time, the "open university" concept was launched. The creation of universities open to all was driven by the need to provide alternative education for adults whose needs could not be met in the traditional classroom. The British Open University began in 1969 through video broadcasting of its weekly courses on the BBC. Over time and with the advent of new technologies, the British Open University's model of distance learning evolved into a student-centered delivery system and administrative structure separate from a campus setting. More economically practical than telecourses, this system envisioned each student as "a node in the network" (Granger, 1990, p. 189) that provides individualized instruction in a virtual classroom. The students have access to a virtual library, customizable based on their particular learning style, and to collaborative tools that encourage discourse and critical thinking (Prewitt, 1998). By encouraging a community of learners, this model overcomes some of the problem of isolation.

During the 1970s, the capability of computers to automate tasks and deliver information made them invaluable tools for many companies, thereby increasing the need for technologically competent workers. This prompted the inception of corporate training programs focused on technology literacy. In schools, word processors, spreadsheets, and database applications enhanced the productivity of educators and students, and the development of educational software offered interactive ways to deliver academic content. As early as 1960s computers were being used with adult learners at the University of Illinois through an integrated learning system called Programmed Logic for Automatic Teaching Operations (PLATO). Integrated learning systems—computer-based instructional programs—were used to distribute educational content through a local network, usually to students in the same room or building. In the 1980s, due to the rapid evolution of information technology and our increasing dependence upon computers as a society, the use of educational technologies expanded considerably. However, during these first few decades, information technologies were used more to automate traditional models of educational delivery than to develop new forms of pedagogy or to enhance learning across distance.

In contrast, during the 1990s widespread usage of the Internet transformed the nature of distance education. As discussed elsewhere in this encyclopedia, the present-day Internet traces its origins to the "ARPANET," a system developed in the late 1960s (Leiner et al., 1997). The ARPANET was initially used by instructors and researchers to share files of information. In 1972, however, e-mail capability was added, transforming computers into a medium that facilitated, direct "people to people" interaction (Weber State University, 1997). In subsequent years, the advancement of networking technologies led to the eventual development of the Internet. Increasing use of personal computers in schools, businesses, and homes helped to establish this budding network of computers. In particular, with the development of the World Wide Web (WWW) as a means of representing and accessing information, Internet use expanded exponentially.

In 1992, it is estimated that the WWW contained a mere 50 Web pages (Maddux, 2001, p. 2). During the 1990s, decreasing prices for computer technologies led to increases in personal ownership, and self-publishing resulted in an explosion of Web-based information. By 2000, the number of Web pages rose to at least one billion (Maddux, 2001). With its capability to facilitate communication between people in various geographic locations and to disseminate information quickly and relatively inexpensively, the Internet appeared well matched for distance education.

During the latter part of the 20th century, the rise of computer-mediated distance education coincided with several important demographic shifts, as well as changes in the world of work. It became less common for a person to spend most of his or her work life in one vocation or one company. Multiple career and job changes over a lifetime were becoming the norm. This increased the need for retraining programs. The citizenry was also becoming increasingly mobile, making location-dependent programs less feasible for many people. Retraining needs led a large number of older adults to seek educational opportunities, but time constraints due to work and family responsibilities made traditional, on-campus programs inaccessible. All of these factors contributed to a growing demand for distance learning programs that businesses and universities hastened to meet.

Beginning in the early 1990s, many universities began using new information technologies both to augment traditional face-to-face classes where instruction occurs in the classroom and to deliver Web-based courses, solely via the Internet. The WWW provided a plethora of resources, which could be accessed inexpensively, around the clock, and from any Internet-accessible location in the world. Course Web sites served as virtual learning "spaces," portals through which course participants gained access to course-related information and reference databases. E-mail facilitated direct, instructor-to-student and student-to-student communication. During the early period of Internet use, the level of information access and communication was unprecedented in distance education.

However, some researchers expressed concern about the efficacy of computer-mediated teaching and learning in contrast with face-to-face instruction (Nickerson, 1993; Peters, 1994). When compared with classroom-based education, some important aspects of human interaction appeared to be lacking. E-mail and Web pages did not capture the spontaneity of face-to-face conversation, the richness of group discussions, and the sense of community prevalent in well-designed classroom settings. The development of computer-mediated communication tools addressed many of these issues.

Asynchronous learning networks (ALNs) based on threaded discussion boards offered a facile way for multiple participants to engage in online conversations and to read and respond to the ideas of their classmates, unrestricted by time. Many students felt that the opportunity to review postings and to think through their own ideas before responding benefited their learning, leading to a higher quality of conversation than they customarily encountered in classroom-based dialogues. Also, the development of Web-based synchronous learning networks (SLNs) added real-time communication to the growing repertoire of distance education tools. In multiuser object-oriented systems (MOOs), multiuser domains (MUDs), and multiuser virtual environments (MUVEs), such as "Tapped In" (http://www.tappedin.org) and "Active worlds" (http://www.activeworlds.com/), learners, instructors, and colleagues were able to gather simultaneously in designated virtual spaces to discuss course-related issues and exchange information of all sorts. ALNs and SLNs proved complementary in addressing different communication preferences and in creating a sense of community among learners.

As technological media have expanded, so has the number of educational institutions and businesses offering instruction via computer and Internet technologies. Computer-mediated learning, which is not time- or location-dependent, provides educational opportunities for those whom traditional schooling might not be feasible, such as working adults, parents of young children, and those with limited mobility. In addition, the high level of autonomy and self-directedness afforded by online education, as well as relative "freedom of pace," has benefited the particular learning needs of adults (Verduin & Clark, 1991).

With the participation of many reputable universities and with increased research documenting effectiveness, distance learning is now a credible and viable mode of instruction. Though detractors exist and accreditation is still a problem for some online programs and virtual universities, the use of information technologies for distance education seems likely to continue growing and evolving, particularly as society invests in research about how to use interactive media to deliver high quality education.

Table 1 presents a timeline of distance learning and new technologies.

DISTANCE EDUCATION AT THE TURN OF THE MILLENNIUM

Socioeconomic Context

Internet-based distance education has arisen amid shifts in the nature of higher education and in socioeconomic conditions worldwide. These changes have contributed to the growth and popularity of the field. The emergence of a global, knowledge-based economy has created new demands on corporate and educational institutions. The rapid evolution of information technologies has served simultaneously to increase those demands and to offer novel ways of developing a highly educated, knowledgeable, and technologically competent citizenry.

According to Selfe (1999, pp. 45–46), growing industrialization throughout the world and "the opening of increasingly competitive global markets" posed a threat to the economic preeminence enjoyed by the U.S. since World War II. In the early 1990s, the U.S. faced rising inflation, high rates of unemployment and poverty, and a widening gap between the rich and poor. Recognizing domestic "high-technology industries [as] the most competitive in the world," the Clinton administration devoted significant resources to the enhancement of the national and global information infrastructures and the domestic computer industry, in order to create new jobs, stimulate income growth, and open new markets for U.S. businesses (Selfe, 1999, p. 51). As information technology became more powerful and less expensive, the U.S. economy hastened its shift from an emphasis on manufactured goods to services and information. In response, the 1995 Economic Report of the President identified the need for increased "investments in education and technology" to prepare a workforce capable of sustaining this new type of economy (Council of Economic Advisors, 1997, p. 7).

Changes in the economy and job market demand new skill sets. The proliferation of knowledge and information has led to "increasingly specialized subfields" (Mitchell & Dertouzos, 1997). This has created a need for workers who possess highly specialized skills and knowledge, often in emerging technology fields. At the same time, the delivery of services and the management of information across wider and more diverse markets requires broader, higher order skills, such as problem solving, project management, and communication. Since individuals today are more mobile and likely to have multiple career and job changes, these types of "soft skills," applicable across working environments, are vital (Thompson, Ganzglass, & Simon, 2000, p. 9).

A multitude of educational programs were created to respond to this demand for new and deeper skills, and corporate training became a thriving segment of the

Table 1 Timeline of Distance Learning and New Technologies

Time Period	Delivery Mechanisms	Benefits	Challenges
Pre–1900	U.S. Postal Service	Rural free delivery allowed students in remote areas to take correspondence courses	Students had to work and learn in isolation
	Railroad	Facilitated speed of mail services and transportation for people and news	Costs of travel and mail limited access
	Traveling Lecturers	Scholars and students could engage in "national" conversation on democracy	Conversation limited to mainly white, upper and middle classes
	Lantern Slides	New technology provided powerful visual aids	Limited to mainly white, upper and middle classes
1900–1960	Radio	Personalized delivery, increased speed	No person-to-person interaction, "one size fits all" content
	Television	Rapid delivery, increased access	No person-to-person interaction, "one size fits all" content
1960–Present	Videotape	Customizable content, reach wide audiences	Costly and complex to produce
	Satellite broadcasting	Wide audiences, real time	Time and place-dependent
	Computers	Rapid delivery of content and communication	Limited face-to-face interaction
	Integrated learning networks	Multiple users, standardized content	No person-to-person interaction, "one size fits all" content
	Internet	Widely accessible, inexpensive and rapid communication across distance	
	World Wide Web	Widely accessible, self-publishing, inexpensive and rapid communication	Limited face-to-face interaction
	Threaded discussion boards	Multiple users, not time dependent	Limited real-time interaction
	MOOs, MUDs, & MUVES	Multiple users, many-to-many communication, rich social interaction, real time	Some bandwidth requirements, limited face-to-face interaction

service industry. Corporations developed retraining programs, and new organizational training businesses formed, offering their services to both companies and individuals. Technology is playing an important role in the delivery of these services, which prepare the workforce for increasingly information-intensive working environments.

Developments in information and communication technologies allowed companies to establish a workforce and serve clientele in a variety of geographical regions. The Internet and Web-based communication tools provided reliable and inexpensive training solutions, to reach learners "in very large numbers, but also in very small classes, or even as individuals, anytime, anywhere" (Thompson, Ganzglass, & Simon, 2000, p. 9). Despite the slowing of the economy in the past couple of years, the need for highly trained, flexible workers remains strong. In an environment of "increased training need and lifelong learning, yet one of diminished resources," the cost effectiveness, accessibility, and adaptability of Internet-based training and education is vital (Schreiber & Berge, n.d.).

Efforts to create more efficient and effective programs, to expand business opportunities, and to recruit new employees led to the development of "corporate universities," which allow a company "to coordinate and manage programs, to train and educate its employees, customers, and suppliers" (Meister, 2001). Corporate universities, which often offer educational programs both to employees and to the general public, rely heavily on Internet technologies to deliver their services and to meet the needs of working adults, who are often constrained by time and location. In addition to creating educational institutions of their own, some businesses are partnering with traditional colleges and universities. This allows them to capitalize on existing pools of learners and infrastructures already in place to educate relatively large numbers of students. For postsecondary institutions, such partnerships create the potential to expand course offerings and career services and receive additional support for distance education initiatives.

Like businesses, colleges and universities have also been impacted by the changing economy, emerging technologies, and the demand for new job skills. These factors have combined with other trends in the field of higher education to expand the need for Internet-based, distance education. Since the end of World War II, colleges and universities have been steadily drawing larger numbers of students from remote locations (Hoxby, n.d.). Enhanced transportation and communication mechanisms have contributed to a population that is increasingly mobile. In addition, the average age of students has risen over time. The once typical 18- to 22-year-old undergrad-in-residence represents less than one-quarter of postsecondary students; colleges and universities are increasingly serving older, working adults, many of whom are part-time students (Greene, 2002).

The shifting economy and the demand for more highly skilled and knowledgeable workers has led to a trend of degree inflation: “85% percent of new jobs require education beyond high school, up from 65 percent in 1991” (Commission on Technology and Adult Learning, 2001, p. 8). This shift has brought many working adults back to school for additional education and retraining. At the same time, the average cost of college tuition “has risen rapidly since 1940,” generally at higher rates than inflation (Hoxby, n.d.).

Characteristics of Distance Education Students

All of these factors have combined to bring students with different needs to the educational market. These students are often working adults who must pay for their own schooling in the face of rising tuition costs and other substantial financial responsibilities. Rather than the traditional, liberal arts education, many seek practical, real-world skills and knowledge that can be directly applied to working environments. Some are highly mobile due to the changing job market and work-related travel, while others have limited mobility due to family responsibilities. Many have considerable time constraints. Internet-based distance education has proven successful in addressing many of the educational needs of this rapidly growing subset of students.

Both the WWW and Internet-based synchronous and asynchronous communication tools allow for the flexible delivery of course content. Learners can access course materials from virtually anywhere in the world, at any time. Unlike traditional correspondence education, students use new technologies to remain in close communication and collaboration with instructors and peers. Since geographic location does not limit participation in virtual learning environments, universities are able to recruit instructors from around the world who would otherwise be unavailable; this brings real-world skills and knowledge into the virtual classroom. In addition, the cost-effectiveness of Internet-based tools makes distance education a viable solution for those who seek education and training in a rapidly evolving job market.

How instructors can know who the students in distance education courses truly are is a topic of concern. Some instructors worry about whether an outside service may be writing the papers some students submit or taking the online tests that courses may require. Web sites exist for students to shop for a paper—for example, term-papers 4u.com promises “original” term papers that will attain good grades, hassle free. Or, one can search at EssayMart.com through over 700,000 term papers. Educators have fought back with their own Web sites; for example, Turnitin.com provides methods for detecting and dealing with plagiarism. As another illustration, Wordcheck is software that detects infringement of intellectual copyright through an extensive database (http://www.wordchecksystems.com/). Debates about how to ensure the validity of student achievements in distance education settings are likely to continue for some time.

Contemporary Delivery Mechanisms for Distance Education

How does one classify a course as distance education? Roberts (1996) suggested four criteria:

1. Teacher and student are separated physically during some portion of the educational experience.
2. There is an overseeing educational organization (this could be a university, secondary school, nonprofit, or even a corporate entity).
3. Educational technology mediates the interaction while apart.
4. Student and teacher can dialogue.

Today’s distance education courses take two forms: courses offered primarily at distance and hybrid courses, also called “distributed learning” courses, which combine face-to-face meetings with virtual experiences (Dede, Whitehouse, & Brown L’Bahy, 2002).

Historically, interactions between teacher and learner were generally unidirectional (e.g., correspondence courses, radio, and television). With the advent of computer-mediated technologies, the emphasis has shifted to interactive tools, allowing for more pedagogically sound instruction. As discussed earlier, computer-mediated communication encompasses both synchronous tools that create a virtual classroom and asynchronous tools that simulate traditional correspondence courses (Guri-Rosenblit, 1999). However, such a categorization is misleading because emerging media enable many novel forms of mediated interaction not well captured by comparisons to prior types of education. Synchronous tools allow students to work together on projects, interact with teachers, and carry on conversations. These tools include Internet-based videoconferencing, multiuser virtual environments, and virtual communities-of-practice. Asynchronous tools permit students to access coursework on their own time and place and allow for deeper reflection than “think-on-your-feet” synchronous tools. Asynchronous media include Internet technologies such as e-mail, the World Wide Web, asynchronous online conferencing systems, learning portals, and “courseware.” Also, some media (such as “groupware” collaboration tools) offer both asynchronous and synchronous options. Groups can meet simultaneously to plan, and later individuals can reenter the virtual space to access and build upon what others have previously done.

Choosing the correct tools for a course or program depends on the aims of the course and the needs of the students. “There is no ‘right’ or ‘wrong’ technology for distance education. Each medium and each technology, through which it can be delivered, has its own strengths and weaknesses" (Moore & Kearsley, 1996, p. 99). The following are illustrative examples of contemporary distance education programs; in each, the tools chosen match the needs of the organization, teacher and learner.

The Fielding Graduate Institute (http://www.fielding.edu/) is an example of a distance learning school. Several programs are offered leading to various accredited degrees. Stevens-Long & Crowell (2002) reported on the development of the Institute's Masters Program in Organizational Management. In this program, an emphasis is placed on using virtual environments, not to replicate face-to-face, but instead to create a space where students can reflect on contributions, interact, and evolve new social constructs of knowledge.

The Network for Education and Technology Services (http://www.unet.maine.edu/) was developed by the University of Maine System to increase participation in post-secondary education (in 1985, 50th in the prevalence of adult education and 47th in the percentage of matriculating high school graduates attending college). Through a tumultuous history, the UNET system has evolved into a support system for colleges within the University of Maine system offering distance education (Carchidi, 2002). One of the highlights of this program is its use of community centers throughout rural Maine to provide a place for students to access equipment and have social interactions.

After identifying potential students and desired outcomes, MCI WorldCom developed an online university (Treanor & Irwin, 2001). With increasing economic pressures and expanding world markets, the corporation's need to have training available at any time and any place increased dramatically. MCI's university utilizes a Web-based portal that allows the company to direct students to appropriate services. MCI WorldCom has tailored its tools, Web-based courses, and streaming video to the needs of the identified student population. Online training programs are discussed elsewhere in this book in more detail.

The distance learning institutions most scrutinized in terms of quality of education are for-profit distance education organizations. These organizations have dramatically increased, with enrollment up 59% in the past 10 years and the market share of four-year programs nearly tripled to 8% of the total. As a group, they are very diverse; the University of Phoenix (http://onl.uophx.edu/) is one of the largest with 75,000 students online across 15 states (Kelley, 2001). Opponents worry that quality and academic standards will become secondary to profits, but proponents suggest that these organizations are more focused on current students' needs than are traditional campuses where individualization is more difficult to achieve than it is online.

Virtual high schools offer distance education programs primarily to secondary school students (elementary programs are just beginning and are still controversial as many people consider social development to be a primary goal of elementary school). One of the pioneers in this area is the "Virtual High School" program originated by the Concord Consortium (http://www.govhs.org/website.nsf). This program currently is enrolling approximately 2000 students in 24 states in a completely asynchronous environment. The students are able to choose from a wide array of courses to enrich their local school's offerings. These courses allow students to experience different perspectives, explore areas of interest in more depth, and possibly escape from stigmas.

One of the drawbacks to distance education has been in finding a suitable replacement for science laboratory experiments. Three options have been explored: "computer simulations, videos of real laboratories and laboratory kits sent to a student" (Forinash & Wisman, 2001, p 40). While these methods decrease costs, the goal of the science lab is to offer students a chance to explore and practice scientific inquiry. Toward this goal, simulations did not allow students much choice, videos offered no options, and safety hazards of unsupervised exploration usually limited home kits. However, several new options for simulations with nearly limitless choices are currently being developed in physics and chemistry to help overcome this deficiency (Carnevale, 2002a, 2002b). Overall, the delivery methods of distance education have come a long way from the initial correspondence courses delivered by horseback. Many students are no longer absorbing information delivered from instructors in isolation, but rather are interacting across distance with teachers and fellow students, questioning, reflecting, and constructing their knowledge. Current distance education programs provide increased access for those with physical and financial limitations. In addition, those with learning styles not met by traditional education can access a wide range of options in distance education programs.

THEORETICAL UNDERSTANDINGS OF THE STRENGTHS AND LIMITS OF CURRENT DISTANCE EDUCATION

The evolution over the past century of research on distance education provides a context for understanding learning across distance. In the early years of scholarship on this topic, typical studies compared student outcomes in traditional face-to-face classrooms with outcomes in correspondence courses. This type of comparative research has sparked a continuing debate among educational researchers, educators, and policymakers, as well as much discussion over the differences between teaching at a distance and teaching in the classroom.

Typically, comparative research methodologies used to analyze face-to-face classrooms and distance learning classrooms reveal a "no significant differences" phenomenon (Russell, 1999).

Russell catalogued 335 comparative studies, and his resulting bibliography was one of the first comprehensive looks at "no significant differences" findings. For example, Crump's 1924 study of learning outcomes between correspondence and classroom students' taking the same courses found "no significant difference" between the two groups. This type of finding continued through the 1940s, when schools began using radio programs for delivery of curriculum (Woelfel & Tyler, 1945), and the 1950s when schools began using closed circuit television for delivery of curriculum (Swartout, 1959). Many present day studies repeat this finding. As illustrations, Linc & Davidson (1995) reported that there was "no significant difference between students using hypertext and those in the traditional classroom," and Navarro and Shoemaker (1999) found that, in six of eight academic variables, there was "no significant difference" between educational outcomes for students learning in the traditional classroom or learning online through computer-mediated technologies. This sort of comparative finding does not reveal much other than that student grades were similar; and, in more recent years, researchers have turned to research designs that reveal more complex and useful findings.

During the 1990s many researchers found troubling flaws in the comparative methodologies used to gauge the effectiveness of online learning. Joy and Garcia (2000) argued "learning effectiveness is a function of effective pedagogical strategies" (p. 33) and should not be measured by delivery systems. Dede, Whitehouse, & Brown-L'Bahy (2002) wrote, "Three decades of research in distance education are largely off-target because studies have typically compared a single medium (such as face-to-face) to another medium (e.g., videoconferencing) for a group. . . . Some students are empowered by each medium, others disenfranchised; the net result is mixed." Comparative studies seemed to tell only a part of the story about the power of interactive media in teaching.

Emerging technologies enable novel applications of theoretical frameworks for teaching and learning that range from constructivist principles of student-centered learning to learning styles based on cultural and affective constructs, as well as learning theories based on distributed cognition and situated learning. As one illustration, Wenger's (1998) research is founded on the assumption that social interactions are the basic process that allows people to learn; this has extended our knowledge of online communities-of-practice. Design research is an example of emerging research methodologies that combine empirical with qualitative research methods based on theories of learning and design. In 1992, design experiments were introduced by Ann Brown and Allan Collins in an effort to use formative analysis as a means to move education research out of the laboratory and into the classroom, as well as to provide a design science for education (Collins et al., in press). To provide robust research results more generalizable than those of comparative research methods, this type of scholarship measures multiple dimensions of design and learner activity in authentic contexts.

The work of Lave and Wenger (1991) in situated learning provides a framework for understanding how computer-mediated knowledge construction shapes learner experience from social, cognitive, and affective perspectives. By using multimedia and other emerging technologies to provide powerful learning experiences that build on the strengths of both distance and classroom learning, distributed learning environments offer multiple entry points for different types of learners (Dede, 1996). Roth's (2001) research in situating cognition grows from an epistemological framework that allows a multidimensional analysis of patterns of student work called "zooming." (Zooming is a research method Roth uses to identify multiple patterns of student behavior in a specific context as it occurs over time. He argues that learners are at different places in their thinking depending on how they relate to the learning environment and that it is important to use a research method that reveals these varying patterns of progress.) Lemke's (2001) analysis of multiple timescale studies of human activity utilized a synthesis of theory and method to allow sophisticated analyses across time and dimensions of behavior; this may reveal more robust findings about how people work and learn. Bielaczyc (2001) has studied developing computer supported collaborative learning communities via focusing on the social infrastructure of these communities and how the design of the tools used to support distance and classroom learning shape learner experiences and community building. Barab et al. (in press) are working to characterize distance learning in terms of the dualities of face-to-face versus online participation in a learning community. Each research method listed seeks to gain more understanding of the challenges and disadvantages of distance education.

DISTANCE/DISTRIBUTED/VIRTUAL LEARNING IN THE FUTURE

"Mixed mode" or "hybrid" courses combine the use of face-to-face teaching with synchronous and asynchronous mediated interaction. "Distributed learning" is a term used to describe such educational experiences, which are distributed across a variety of geographic settings, time, and various interactive media (Dede, Brown-L'Bahy, & Whitehouse, 2002). Within a decade, "distance education" may be an obsolete concept, as may the term "face-to-face education." Instead, all instruction may be some balance between classroom-based and distance-based learning interactions, determined by the subject matter, student population, and educational objectives. Emerging interactive media are facilitating such an evolution.

The Evolution of Interactive Media

Table 2 lists devices, media, and virtual contexts supported by sophisticated information technologies, along with the estimates of a conservative timeframe for their technological and economic feasibility (Dede, 2000). Note that many current capabilities are not yet widely used instructionally. Researchers rapidly employ new knowledge-sharing media, but pedagogical and assessment practices in schools, campuses, and organizational training settings are slow to alter.

The important issue for distance education is not the availability and affordability of sophisticated computers and telecommunications, but the ways these devices enable powerful learning situations that aid students in extracting meaning out of complexity. New forms of representation (e.g., interactive models that utilize visualization and other means of making abstractions tangible and sensory) make possible a broader, more powerful repertoire of pedagogical strategies. Also, emerging interactive media empower novel types of learning experiences; for example, interpersonal interactions across networks can lead to the formation of virtual communities.

Table 2 Devices, Media, and Virtual Contexts Grouped by Timeframe.

Functionality	Uses	Time Frame
Cognitive audit trails (automatic recording of user actions)	Support for finding patterns of suboptimal performance	Current
Intelligent tutors and coaches for restricted domains	Models of embedded expertise for greater individualization	Current
User-specific, limited-vocabulary voice recognition	Restricted natural language input	Current
High quality voice synthesis	Auditory natural language output	Current
Fusion of computers, telecommunications	Easy interconnection; universal "information appliances"	2005
Information "utilities" (synthesis of media, databases, and communications)	Access to integrated sources of data and tools for assimilation	2005
Microworlds (limited, alternate realities with user control over rules)	Experience in applying theoretical information in practical situations	2005
Semi-intelligent computational agents embedded in applications	Support for user-defined independent actions	2008
Advanced manipulatory input devices (e.g., gesture gloves with tactile feedback)	Mimetic learning which builds on real-world experience	2008
Artificial realities (immersive, multisensory virtual worlds)	Intensely motivating simulation and virtual experience	2010
"Information appliance" performance equivalent to current supercomputers	Sufficient power for simultaneous advanced functionalities	2010
Consciousness sensors (input of user biofeedback into computer)	Monitoring of mood, state of mind	2010
Artifacts with embedded semi-intelligence and wireless interconnections	Inclusion of "smart devices" in real-world settings	2010

Over the next decade, three complementary interfaces will shape how people learn (Dede, 2002):

- *The familiar "world to the desktop" interface*, providing access to distant experts and archives, enabling collaborations, mentoring relationships, and virtual communities-of-practice. This interface is evolving through initiatives such as Internet2.
- *Interfaces for "ubiquitous computing,"* in which portable wireless devices infuse virtual resources as we move through the real world. The early stages of "augmented reality" interfaces are characterized by research on the role of "smart objects" and "intelligent contexts" in learning and doing.
- *"Alice-in-Wonderland" multiuser virtual environment interfaces*, in which participants' avatars interact with computer-based agents and digital artifacts in virtual contexts. The initial stages of studies on shared virtual environments are characterized by advances in Internet games and work in virtual reality.

The following vignette illustrates how applying ubiquitous computing to embed virtual learning throughout the real world might reshape the nature of education.

> Alec and Arielle strolled through Harvard Yard on their way to the museum, to collect data for their class assignment. Each carried a handheld device (HD) that softly pulsed every time they walked past a building in the Yard. The vibration signaled that the building would share information about its architecture, history, purpose, and inhabitants, using interactive wireless data transfer. Sometimes Alec would stop and use his HD to ask questions about an interesting looking location. Today, he was in a hurry and ignored the pulses.
>
> Inside the museum, Alec and Arielle split up to work on their individual assignments. When Alec typed his research topic into the museum computer, it loaded a building map into his HD, with flashing icons showing exhibits on that subject. At each exhibit, Alec could capture a digital image on his HD, download data about the artifacts and links to related Web sites, and access alternative interpretations about the exhibit. His HD automatically supplied information about Alec's age and background to ensure that the material he received was appropriate in native language, reading level, and learning style. While the museum-supplied information was interesting, Alec always enjoyed the comments posted about each exhibit by other kids. Sometimes, he added a few remarks of his own to the ongoing discussion. Seeing a cool artifact related to Arielle's topic, Alec paused to link to her HD, sending a digital image of the exhibit and information on its location.

> Alec's favorite exhibits were those augmented by virtual environments. For example, at a panorama showing the bones found at a tar pit, Alec's HD depicted a virtual reconstruction of the dinosaurs that were trapped at that prehistoric location. In the virtual environment, he could assume the perspective of each species and walk or fly or swim through its typical habitat. Other types of exhibit-linked virtual environments enabled "time travel" to show how a particular spot on the Earth's surface had changed over the eons. For each epoch, Alec used virtual probes on his HD to collect data about temperature, air pressure, elevation, and pollutants.
>
> Walking back from the museum, Arielle and Alec shared what they had found. Both wondered what learning was like before augmented reality and ubiquitous computing, when objects and locations were mute and inert. How lifeless the world must have been!

This vision illustrates how richly interwoven real and virtual educational experiences may become as distance learning media continue to evolve.

CONCLUSION

In the long run, distributed learning can potentially conserve scarce financial resources by maximizing the educational usage of information devices (televisions, computers, telephones, video games) in homes and workplaces. In addition, distributed learning enables shifts in the pattern of society's investments in education. Less money is needed for physical infrastructure—buildings, parking lots—and more resources can go into ways of creating a virtual community for creating, sharing, and mastering knowledge.

However, in education's future, keeping a balance between virtual interaction and direct interchange is important (Dede, 1996). Technology-mediated communication and experience supplement, but do not replace, immediate involvement in real settings; thoughtful and caring participation is vital for making these new capabilities truly valuable in complementing face-to-face interactions. How a medium shapes its users, as well as its message, is a central issue in understanding the transformation of distance education into distributed learning. The telephone creates conversationalists; the book develops "imaginers," who can conjure a rich mental image from sparse symbols on a printed page. Much of television programming induces passive observers; other shows, such as Sesame Street and public affairs programs, can spark users' enthusiasm and enrich their perspectives. As we move beyond naive "information superhighway" concepts to envision the potential impacts of learning across distance, society will face powerful new interactive media capable not only of great good, but also misuse. The most significant influence on the evolution of virtual learning will not be the technical development of more powerful devices, but the professional development of wise designers, educators, and learners.

GLOSSARY

Augmented reality Real-world situations into which virtual representations are infused or superimposed.

Avatar A character within a virtual environment that is controlled directly by the user and interacts with the environment as the user's proxy.

Chat Online (Web-based), text-based, real-time communication tool.

Community of learners A group of learners, working collaboratively toward shared, educational goals.

Computational agent A character built into a software application that responds to users' actions in ways determined by its designer.

Constructivist A theory of learning based on students' building new knowledge through interpreting experiences; philosophies about constructivism have varying positions about how much instructional guidance is involved in this process.

Design experiments An emerging research method that builds theory through iterative cycles of testing in real-world settings.

Distance education/distance learning Teaching and learning across distance and time.

Distributed cognition Dispersal of intellectual functioning across physical, social, and symbolic supports.

E-learning Learning and teaching via Internet technologies.

Intelligent context An environment that responds to users' actions in ways determined by its designer.

Lyceum lecturers Intellectuals who traveled the United States over a century ago, bringing intellectual debate about democracy to everyday Americans.

Lantern slides An early version of projected slides or overheads.

Listening groups Small communities of people who gathered to listen to radio broadcasts, followed by discussions on the content.

Microworld A computer-based environment whose dynamics are based on causal factors that the user can manipulate.

Mimetic learning Learning based on copying the actions of an expert.

Multiuser virtual environment A computer application that simulates a context in which multiple participants can interact and experience simulated pheno mena.

Situated cognition The theory that knowledge is rooted in the context in which it is learned.

Situated learning A theory of learning based on students' acquiring knowledge in an authentic context with a community of participants that range from novices to experts.

Smart objects Technological tools that respond to a user's activity in ways determined by the designer.

Ubiquitous computing The infusion of smart objects and intelligent contexts throughout a real-world setting.

Web-enhanced (courses) Traditional, face-to-face cour ses that incorporate some Internet-based learning and teaching experiences.

Virtual community of practice People with similar roles who collectively use mediated interaction to share ideas about accomplishing those roles.

Virtual reality A computer interface based on multisensory immersion.

CROSS REFERENCES

See *Digital Libraries; Intelligent Agents; Internet Literacy; Online Communities; Research on the Internet; Web-Based Training.*

REFERENCES

Banas, E. J., & Emory, W. F. (1998). History and issues of distance learning. *Public Administration Quarterly, 22*(3), 365–383.

Barab, S., MaKinster, J., & Scheckler, R. (in press). Designing system dualities: Characterizing online community. In S. Barab, R. Kling, & J. Gray (Eds.), *Designing for virtual communities in the service of learning*. Cambridge, MA, Cambridge University Press.

Bielaczyc, K. (2001, March). Designing social infrastructure: The challenge of building computer-supported learning communities. Paper presented at *Euro-CSCL 2001 Conference,* Maastricht, The Netherlands.

Bourke Distance Education Centre (2002, October 31). *Bourke School of Distance Education*. Retrieved October 31, 2002, from Bourke Distance Education Centre Web site at http://www.bourkep-d.schools.nsw.edu.au/index.htm

Bunker, E. L. (1998). Gaining perspective for the future of distance education from early leaders. *American Journal of Distance Education, 12*(2), 46–53.

Carchidi, D. M. (2002). *The virtual delivery and virtual organization of postsecondary education*. New York: Routledge Falmer.

Carnevale, D. (2002a, December 10). Online-lab software simulates chemical interactions—And explosions. *The Chronicle of Higher Education*. Retrieved December 18, 2002, from http://chronicle.com/free/2002/12/2002121001t.htm

Carnevale, D. (2002b, December 16). A virtual laboratory simulates physics experiments. *The Chronicle of Higher Education*. Retrieved December 18, 2002, from http://chronicle.com/free/2002/12/2002121601t.htm

Collins, A., Joseph, D., & Bielaczyc, K. (in press) Design research: Theoretical and methodological issues. *Journal of the Learning Sciences*, special issue.

Commission on Technology and Adult Learning (2001). *A vision of e-learning for America's workforce*. Retrieved November 11, 2002, from National Governors Association Web site at http://www.nga.org/cda/files/ELEARNINGREPORT.pdf

Council of Economic Advisors (1997). *Economic report of the President*. Retrieved November 11, 2002, from United States Government Web site at http://w3.access.gpo.gov/usbudget/fy1998/pdf/erp.pdf

Crump, R. E. (1924). Correspondence and class extension work in Oklahoma. In *Education* (pp 31–43). New York: Teachers College, Columbia University.

Cuban, L. (2001). *Oversold and underused computers in the classroom*. Cambridge, MA: Harvard University Press.

Dede, C. (1996). Emerging technologies and distributed learning. *American Journal of Distance Education, 10*(2), 4–36.

Dede, C. (2000). Emerging technologies and distributed learning in higher education. In D. Hanna (Ed.), *Higher education in an era of digital competition: Choices and challenges* (pp. 71–92). New York: Atwood.

Dede, C. (2002). Vignettes about the future of learning technologies. In *2020 Visions: Transforming Education and Training through Advanced Technologies*. Washington, DC: U.S. Department of Commerce. Retrieved January 20, 2003, from http://www.ta.doc.gov/reports/TechPolicy/2020Visions.pdf

Dede, C., Whitehouse P., & Brown-L'Bahy, T. (2002). Designing and studying learning experiences that use multiple interactive media to bridge distance and time. In C. Vrasidas & G. V. Glass (Eds.), *Distance education and distributed learning* (pp. 1–29). Greenwich, CT: Information Age Publishing.

Dymock, D. (1995). Learning in the trenches: A World War II distance education. *Distance Education, 16*(1), 107–119.

Field, P. S. (2001). "The transformation of genius into practical power": Ralph Waldo Emerson and the public lecture. *Journal of the Early Republic, 21(3)*, 467–493.

Forinash, K., & Wisman, R. (2001). The viability of distance education science laboratories. *T.H.E. Journal, 29*(2), 38–45.

Froke, M. D. (1995). Antecedents to distance education and continuing education: Time to fix them. *New Directions for Adult and Continuing Education, 67*, 61–70.

Funk, C. (1998). The Art in America radio programs, 1934–1935. *Studies in Art Education, 40*(1), 31–45.

Granger, D. (1990). Open universities. *Change, 22*(4), 44.

Greene, C. (2002). *Interview*. Retrieved November 11, 2002, from Selling to Schools Web site at http://www.sellingtoschools.com/interviews/kgreen.html

Guri-Rosenblit, S. (1999). *Distance and campus universities: Tensions and interactions; a comparative study of five countries*. New York: IAU Press.

Hoxby, C. M. (n.d.). *The effects of geographic integration and increasing competition in the market for college education*. Retrieved November 11, 2002, from http://post.economics.harvard.edu/faculty/hoxby/papers/exp_tuit.pdf

Jeffries, M. (2002, October 31). *IPSE—Research in distance education*. Retrieved October 31, 2002, from IHETS Web site at http://www.ihets.org/consortium/ipse/fdhandbook/resrch.html

Joy, E., & Garcia, F. E. (2000). Measuring learning effectiveness: A new look at no-significant-difference findings. *Journal of Asynchronous Learning Networks, 4*(1), 22–37.

Kelley, K. F. (2001, July). *Meeting needs and making profits: The rise of for-profit degree-granting institutions*. Retrieved November 24, 2002, from http://www.ecs.org/clearinghouse/27/33/2733.htm

Lave, J., & Wenger, E. (1991). *Situated learning legitimate peripheral participation*. Cambridge, UK: Cambridge Univ. Press.

Leiner, B. M., Cerf, V. G., Clark, D. D., Kahn, R. E., Kleinrock, L., Lynch, D. C., et al. (1997). *A brief* *history of the Internet*. Retrieved October 28, 2002, from http://www.mathcs.emory.edu/~cheung/history-internet.html

Lemke, J. L. (2001). The long and the short of it: Comments on multiple timescale studies of human activity. *The Journal of the Learning Sciences, 10(1&2)*, 17–26.

Linc, C., & Davidson, G. (1995). Effects of linking structure and cognitive style on students' performance and attitude in a computer-based hypertext environment. Paper presented at the National Convention of the AECT, Nashville, TN.

Maddux, C. D. (2001). *Educational computing: Learning with tomorrow's technologies*. Needham Heights, MA: Allyn & Bacon.

Meister, J. C. (2001, February 9). The brave new world of corporate education. *The Chronicle of Higher Education*. Retrieved November 13, 2002, from http://chronicle.com/free/v47/i22/22b01001.htm

Mitchell, W. J., & Dertouzos, M. L. (1997). *Educational Technology Council Report, 1997*. Cambridge, MA: Masschusetts Institute of Technology. Retrieved November 11, 2002, from http://web.mit.edu/committees/councils/edtech/Ed_Tech_Chap3.HTM

Mood, T. A. (1995). *Distance education, An annotated bibliography*. Englewood, CO: Libraries Unlimited.

Moore, M. G., & Kearsley, G. (1996). *Distance education: A systems view*. Belmont: Wadsworth.

Moran, L. (1993). Genesis of the Open Learning Institute of British Columbia. *Journal of Distance Education, 8*(1), 43–70.

Nasseh, B. (1997). *A brief history of distance education*. Retrieved October 13, 2002, from Senior Net Web site: http://www.seniornet.org/edu/art/history.html

Navarro, P., & Shoemaker, J. (1999). The power of cyberlearning: An empirical test. *Journal of Computing in Higher Education, 11(1)*. 29–54.

Nickerson, R. S. (1993). On the distribution of cognition: Some reflections. In G. Salomon (Ed.), *Distributed cognitions: Psychological and educational considerations* (pp. 87–102). Cambridge, UK: Cambridge Univ. Press.

PBS (2002, October 31). *Distance learning, An overview*. Retrieved October 15, 2002, from PBS Web site at http://www.pbs.org/als/dlweek/index.html

Peters, O. (1994). Distance education and industrial production: A comparative interpretation in outline (1967). In D. Keegan (Ed.), *The industrialization of teaching and learning* (pp. 107–127). London, UK: Routledge.

Prewitt, T. (1998). The development of the distance learning delivery system. *Higher Education in Europe, 23*(2), 187–194.

Roberts, J. M. (1996). The story of distance education: A practitioner's perspective. *Journal of the American Society for Information Science, 47*(11), 811–816.

Roth, W.-M. (2001). Situating cognition. *The Journal of the Learning Sciences, 10*(1&2), 27–61.

Russell, T. L. (1999). The "No Significant Difference Phenomenon" online source. Retrieved November 8, 2002, from TeleEducation NB Web site at http://teleducation.nb.ca/nosignificantdifference

Schreiber, D. A., & Berge, Z. L. (n.d.). *Distance training: How to use organizational technology to implement distance and distributed learning*. Retrieved November 11, 2002, from http://userpages.umbc.edu/~berge/introdt.htm

Scott, J. C. (1999). The Chautauqua Movement: Revolution in popular higher education. *The Journal of Higher Education, 70*(4), 389–412.

Selfe, C. L. (1999). *Technology and literacy in the twenty-first century*. Carbondale, IL: Southern Illinois Univ. Press.

Shea, R. H., & Boser, U. (2001, October 15). Special report: E-learning guide. *U.S. News & World Report, 131*, 44.

Stevens-Long, J., & Crowell, C. (2002). *The design and delivery of interactive online graduate education. Handbook of Online Learning*. Thousand Oaks, CA: Sage.

Sumner, J. (2000). Serving the system: A critical history of distance education. *Open Learning, 15*(3), 267–285.

Swartout, S. G. (1959). *Report on findings in educational television*. Washington, D.C.: NEA Department of Visual Instruction.

Tait, A. (1994). The end of innocence: Critical approaches to open and distance learning. *Open Learning, 9*(3), 27–36.

Thompson, C., Ganzglass, E., & Simon, M. (2000). *The state of e-learning in the states*. Retrieved November 11, 2002, from National Governors Association Web site: http://www.nga.org/cda/files/060601ELEARNING.pdf

Treanor, C., & Irwin, J. P. (2001). The world is officially open for business. In Z. L. Berge (Ed.), *Sustaining distance training* (pp. 70–84). San Francisco: Jossey-Bass.

Tyack, D. B. (Ed.). (1967). *Turning points in American educational history* (Blaisdell book in education). Waltham, MA: Blaisdell.

Verduin, J. R., & Clark, T. A. (1991). *Distance education: The foundations of effective practice*. San Francisco: Jossey-Bass.

Weber State University (1997). *A brief history of the Internet*. Retrieved October 38, 2002, from Weber State University Web site at http://www.chpweb.weber.edu/hthsci/apio/history/default.html

Wenger, E. (1998). *Communities of practice*. Cambridge, UK: Cambridge Univ. Press.

Woelfel, N., & Tyler, I. K. (1945). *Radio and the school*. Tarrytown-on-Hudson, NY: World Book.

FURTHER READING

Fleischman, J. (n.d.). *Distance learning and adult basic education*. Retrieved January 17, 2003, from http://www-tcall.tamu.edu/hopey/09.pdf

Malan, R. F., Rigby, D. S., & Glines, L. J. (1991). Support services for the independent study student. In B. L. Watkins & S. J. Wright (Eds.), *The foundations of american distance education: A century of collegiate correspondence study* (pp. 159–172). Dubuque, IA: Kendall/Hunt.

Matthews, D. (1999). The origins of distance education and its use in the United States. *T.H.E. 27*(2), 54–67.

Downloading from the Internet

Kuber Maharjan, *Purdue University*

INTRODUCTION

The Internet is becoming increasingly popular. Millions of people use this technology to communicate through e-mail, conduct research and distance learning, and engage in Internet chat, teleconferencing, video conferencing, and document conferencing. They use Internet technology to exchange information via the Web. To use this technology successfully, they must understand the fundamentals of downloading from the Web.

Downloading means getting information *(files)* to your local computer *(usually a client)* from a remote computer *(usually a server)*. Conversely, *uploading* means transferring files from your local computer to a remote computer (Figure 1).

Hereafter, the term "downloading" refers to the entire process from selecting the file (or files) to be downloaded to receiving the file (or files) on the receiving computer.

Why would users want to download from the Internet? Potential users can listen to the music they like, read news or magazines, and watch movies. Because the file size of audio and video is usually very large, these types of files are usually streamed. If users download a file from the Internet this downloading results in a persistent copy residing on the user's PC. On the other hand, streamed audio or video results in a nonpersistent version. As will be discussed later, users can download files using file transfer protocol (FTP) or hypertext transfer protocol (HTTP). All downloaded materials are usually the intellectual property of the author (or authors). Therefore, users may be required to agree to their terms and conditions stipulated by the owner. Some material available for downloading may be illegal, for example, child pornography and hate literature.

Typically, users would download from the Internet to get new applications or documents and update the current ones.

DATA COMPRESSION FOR FASTER DOWNLOADS

Compression refers to converting a file to a smaller file. There are many ways to compress files to different formats without losing too much detail (images, audio, and video). Some compression techniques are lossy and some are not. In lossy data compression techniques, redundant or unnecessary information is removed, and some amount of data is lost. This enables computers and other devices to transmit the same amount of information in fewer bits. Some examples are converting from BMP or TIFF to JPG or GIF, WAV to MP3 or WMA. For example, a 100-MB (Megabyte => 2^{20} => 1,048,575 bytes) .bmp file, may be converted to a 3-MB .jpg file. Other technologies such as WinZip, UltimateZip, PKZip, JAR, and Interface de Compression Ergonomique pour Windows (ICEOWS) help compress multiple files and folders into one file. Different types of files compress to different degrees, and some of the different compression tools vary in the extent to which they can compress a file. Most text and audio/video files are highly compressible whereas program files are not. This compressed file is smaller in size and more quickly and easily downloaded. Downloaded files can be restored to the original files and folders after these technologies have been applied.

Most of these utilities allow users to view all the file names, folder names, and subfolder names, and they allow all or only selected files and/or only necessary files and/or folders to be extracted. Moreover, these technologies have other advantages: WinZip is very popular because it is easy to use; UltimateZip, ICEOWS, and JAR are free.

Connections between home computers and the Internet can vary widely in speed. Conventional dial-up modems deliver up to 56 kilobits per second (Kbps). Kbps is a transfer speed where as baud is the number of signaling events per second. In this case the bandwidth (the connection speed between a client machine and a server machine) is 56 Kbps. A dedicated phone connection such as a T-1 delivers up to 1.544 Mbps (Megabits per second, 1 Mbps = 1,000,000 bits per second). Users need to be aware of the effect their connection speed will have on the time needed to download a file. For example, a 25-MB file will take just over an hour (62 minute) at 56 Kbps, but less than 2 minute with a typical cable modem. The same 25-MB file, if compressed to 3 MB, reduces the transfer times to 7½ minute for the dial-up modem and around

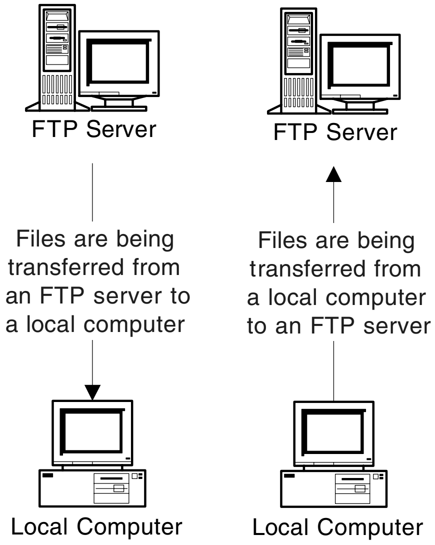

FTP Server
FTP Server
Files are being transferred from an FTP server to a local computer
Files are being transferred from a local computer to an FTP server
Local Computer
Local Computer

Figure 1: Downloading and uploading.

15 second for the cable modem. Therefore, it is important to check the speed of the Internet connection and size of the file before it is downloaded. Then individuals can calculate the approximate time for downloading. Usually, once the downloading process is started, most processes will display approximate time left to download, the average transfer rate, the percentage of the completion of download, and the size of the file already transferred.

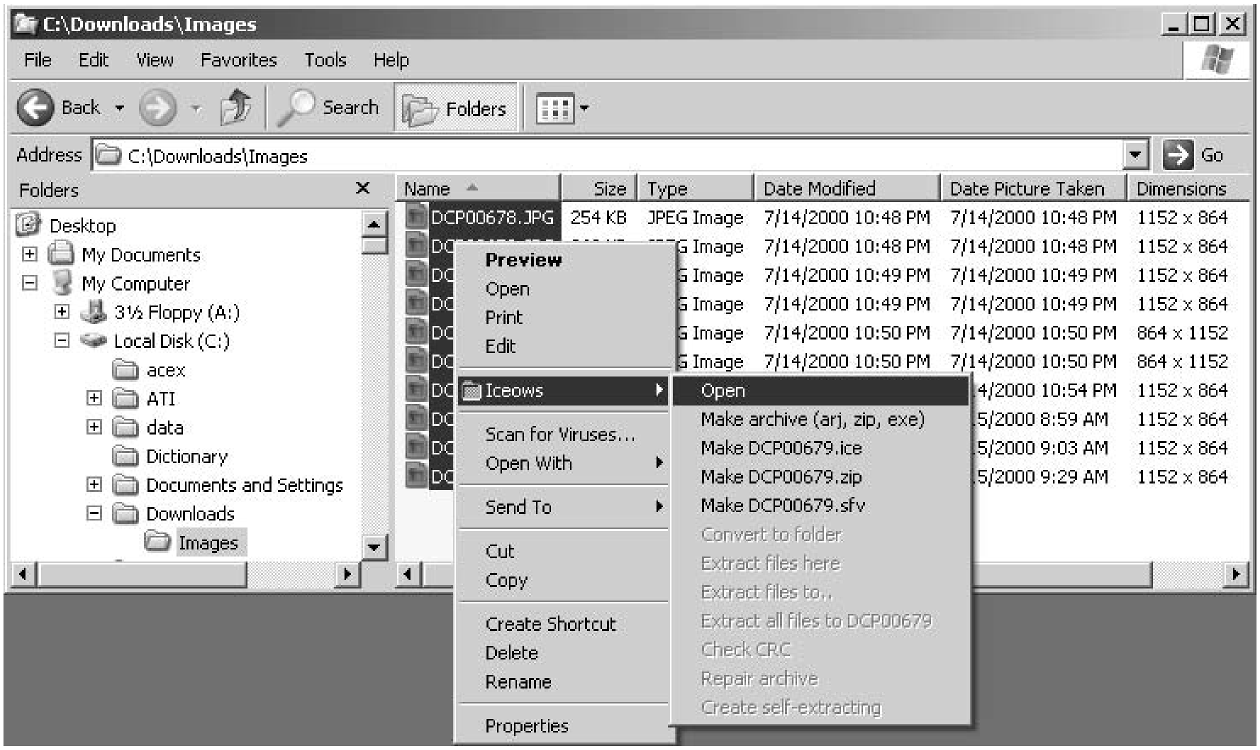
How are compressed files produced? Figure 2 illustrates creating a ZIP (compressed) file using the ICEOWS program. The author has chosen this program rather than the more popular WinZip because this program is free, efficient, and easy to use; it also supports a variety of compression formats. The same program can be used to uncompress the downloaded compressed files.

This program can be downloaded from http://www.iceows.com and installed following the instructions described in Downloading Steps. Similar programs for Macintosh (e.g., MacZip and SmartZip) and Linux (e.g., GnomeZip and kArchiveur) can be downloaded from http://tucows.com (TuCows, n.d.).

> ICEOWS is written to compress or extract easier archive files. With ICEOWS you can extract, test, read properties and comments or view files stored in ICE, ARJ, ZIP, GunZip, TAR, Microsoft CAB, RAR, ACE, MIME, Mac HQX, UUEncode, XXEncode, Base64, JAR, EAR, WAR, LHA, IMP, BZ2 files without external program. If the archive is an ICE, ARJ, or ZIP, you also can add, update, delete files stored in the archive. It integrates all of these functions in the same interface that the Windows 9x, Me, NT or XP explorer use. (http://www.iceows.com/HomePageUS.html)

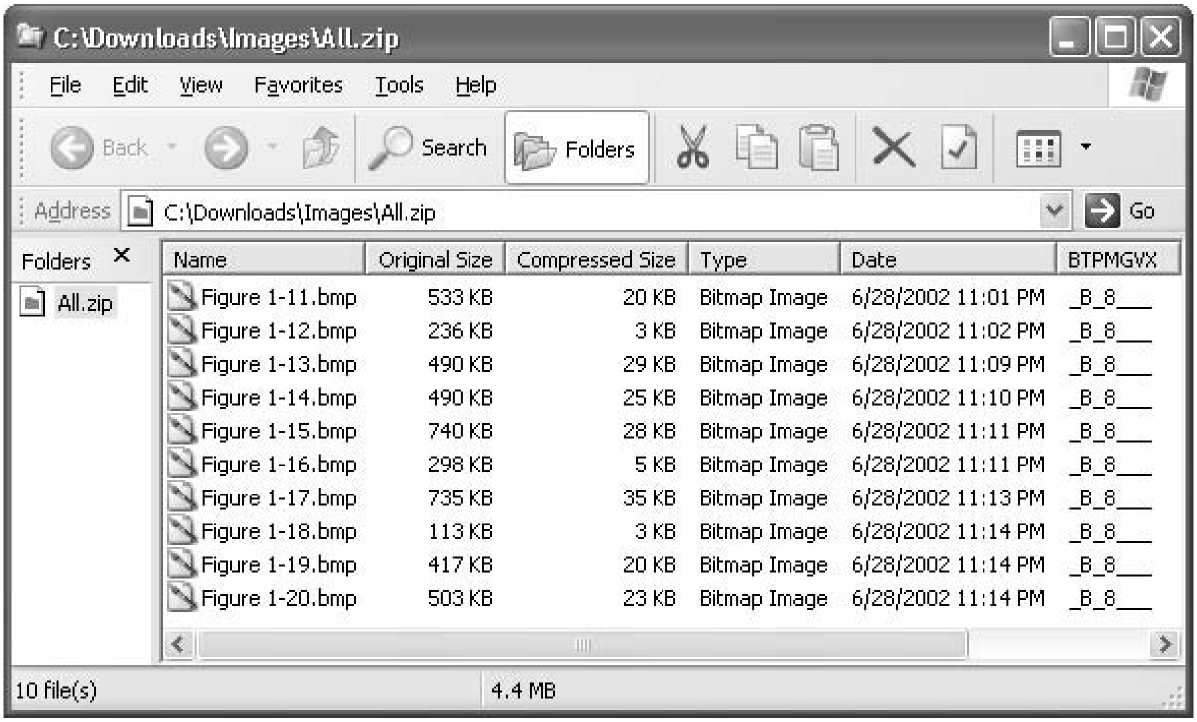
Figure 3 depicts the contents of All.zip in Windows XP Explorer. As can be seen from the original sizes and compressed sizes in Figure 3, 10 BMP pictures of 4.4 MB are compressed to a 188-KB (All.zip) file. In this particular case, the compression is over 95%. This translates to a reduction in the download time of 95%.

Because data compression increases storage capacity tremendously, this technology is also widely used in backup utilities, database management systems, and other applications. While users may transfer the files in



Name	Size	Type	Date Modified	Date Picture Taken	Dimensions
DCP00678.JPG	254 KB	JPEG Image	7/14/2000 10:48 PM	7/14/2000 10:48 PM	1152 x 864
DC		G Image	7/14/2000 10:48 PM	7/14/2000 10:48 PM	1152 x 864
DC		G Image	7/14/2000 10:49 PM	7/14/2000 10:49 PM	1152 x 864
DC		G Image	7/14/2000 10:49 PM	7/14/2000 10:49 PM	1152 x 864
DC		G Image	7/14/2000 10:50 PM	7/14/2000 10:50 PM	864 x 1152
DC		G Image	7/14/2000 10:50 PM	7/14/2000 10:50 PM	864 x 1152
DC				4/2000 10:54 PM	1152 x 864
DC				5/2000 8:59 AM	1152 x 864
DC				5/2000 9:03 AM	1152 x 864
DC				5/2000 9:29 AM	1152 x 864

Figure 2: Creating a ZIP (compressed) file using ICEOWS program.



Name	Original Size	Compressed Size	Type	Date	BTPMGVX
Figure 1-11.bmp	533 KB	20 KB	Bitmap Image	6/28/2002 11:01 PM	_B_8___
Figure 1-12.bmp	236 KB	3 KB	Bitmap Image	6/28/2002 11:02 PM	_B_8___
Figure 1-13.bmp	490 KB	29 KB	Bitmap Image	6/28/2002 11:09 PM	_B_8___
Figure 1-14.bmp	490 KB	25 KB	Bitmap Image	6/28/2002 11:10 PM	_B_8___
Figure 1-15.bmp	740 KB	28 KB	Bitmap Image	6/28/2002 11:11 PM	_B_8___
Figure 1-16.bmp	298 KB	5 KB	Bitmap Image	6/28/2002 11:11 PM	_B_8___
Figure 1-17.bmp	735 KB	35 KB	Bitmap Image	6/28/2002 11:13 PM	_B_8___
Figure 1-18.bmp	113 KB	3 KB	Bitmap Image	6/28/2002 11:14 PM	_B_8___
Figure 1-19.bmp	417 KB	20 KB	Bitmap Image	6/28/2002 11:14 PM	_B_8___
Figure 1-20.bmp	503 KB	23 KB	Bitmap Image	6/28/2002 11:14 PM	_B_8___

Figure 3: Data compression example.

compressed form, it is almost certain that any application that uses the file(s) as data or as an application requires those files to be uncompressed and to remain in their full, uncompressed size. This has consequences for disk consumption. Disk usage will be based on the original, uncompressed sizes, not the compressed sizes.

SAVING AND ORGANIZING DOWNLOADED FILES/FOLDERS

When downloading applications and updates, users who do not remember the name of the original files or who have not renamed or stored the files properly may find it impossible to retrieve a downloaded file. Therefore, it is imperative that they organize their downloaded files and folders carefully in order to find them quickly and easily. When renaming a file, they should use descriptive words, so that it is easier to understand the file without uncompressing or installing software. Precise renaming will help them understand the type of the file by looking at the downloaded file name at any time.

Before downloading, individuals should verify that they have enough room to store the downloaded files. Even though users can organize the directory structure for downloaded files/folders in any fashion, they should organize them in such a way that it is easier to locate the files that they have downloaded. An example of a directory structure is shown in Figure 4.

With the introduction of the newer versions of Web browsers (Internet Explorer and Netscape Navigator), it is possible, in some cases, to open a Web site as a Web folder (Figure 5). Using this newer technology, users can treat a Web site folder as a regular folder. Once the successful connection has been made with proper authorization, users can upload/download from this site as if they were transferring files to and from local storage devices. To transfer files, users can just drag and drop. Newer folders and files can be created directly on the Web site as if users were creating files and folders in their local storage devices. Connection speed clearly determines the time necessary to accomplish download and upload tasks.

Figure 6 shows how to initiate opening a Web site as a Web folder under Internet Explorer. Netscape Navigator also supports a Web folder. After typing the URL of the Web site or selecting from the list box, the “Open as Web Folder” check box must be selected to open the Web site as a Web folder. If the particular Web site is password protected, a log-in dialog box will be presented (Figure 7). After the correct user name and the password have been entered, users will have access to Web folders as if those folders were part of their local storage (hard disk). If the Web site is not password protected, the Web folder screen will be presented automatically. Figure 8 depicts a Web folder.

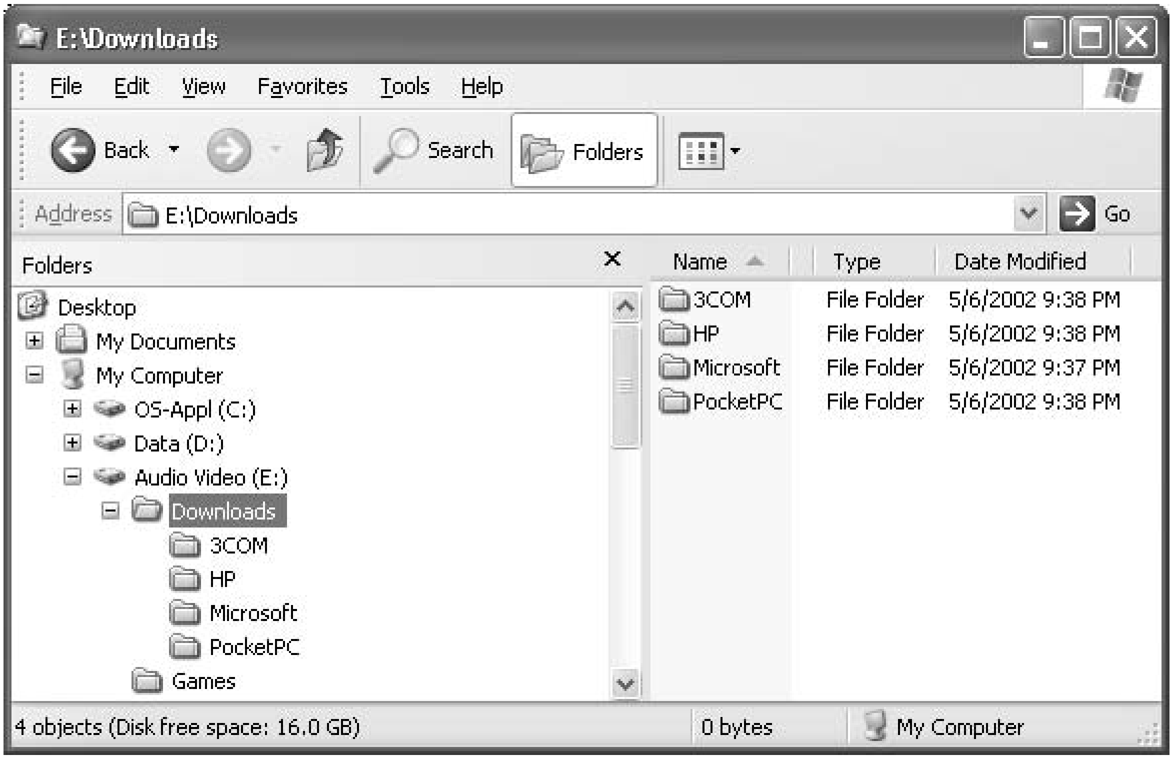
Depending upon the types of access rights assigned to Web folders and subfolders under the Web folders to specific users, authenticated users may be able to perform the following tasks:

Downloading files and folders from the Web folders—in this case, the users drag and drop as if using Windows Explorer.

Uploading files and folders to the Web folders—users also drag and drop.

Creating new folders—users right click and create a new folder.

Removing existing files and folders.

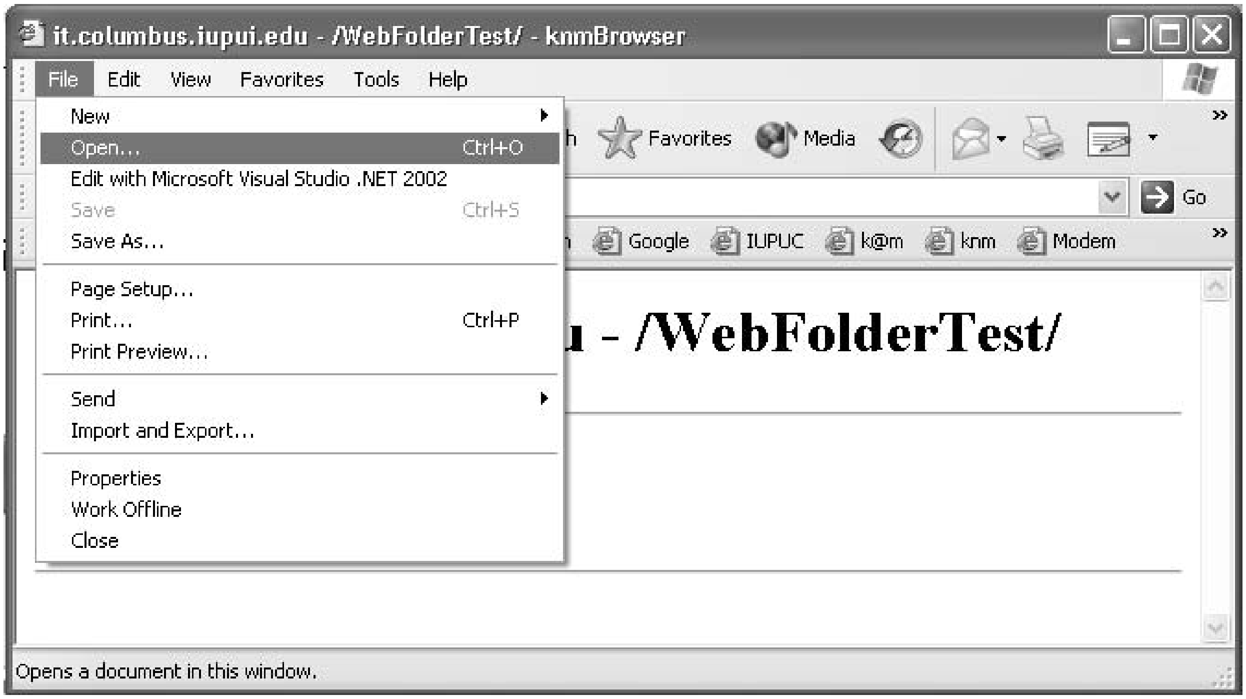


Name	Type	Date Modified
3COM	File Folder	5/6/2002 9:38 PM
HP	File Folder	5/6/2002 9:38 PM
Microsoft	File Folder	5/6/2002 9:37 PM
PocketPC	File Folder	5/6/2002 9:38 PM

4 objects (Disk free space: 16.0 GB) 0 bytes My Computer

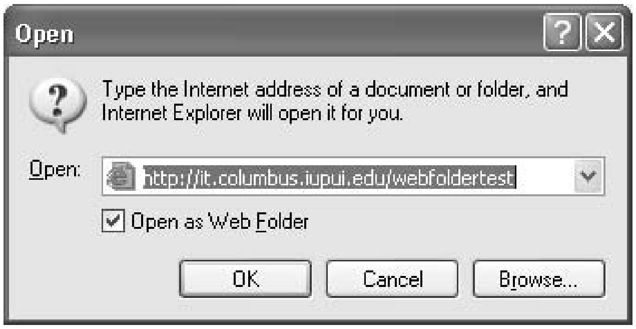
Figure 4: Organizing downloaded files and folders.

In Figure 8 individuals can see a Web folder screen under Windows XP. The author has created a new folder in a Web folder and renamed it to "CreatedFromWeb." This form of downloading and uploading greatly simplifies transferring files from the Internet to the users. Once the Web folder is opened, downloading and uploading are as simple as copying and moving files and folders around local hard drives.



Opens a document in this window.

Figure 5: Opening a Web site as a Web folder, similar to opening a Web site.



Open

Type the Internet address of a document or folder, and Internet Explorer will open it for you.

Open: http://it.columbus.iupui.edu/webfoldertest

☑ Open as Web Folder

Figure 6: Opening a Web site as a Web folder.

DOWNLOADING SPEED AND TIME

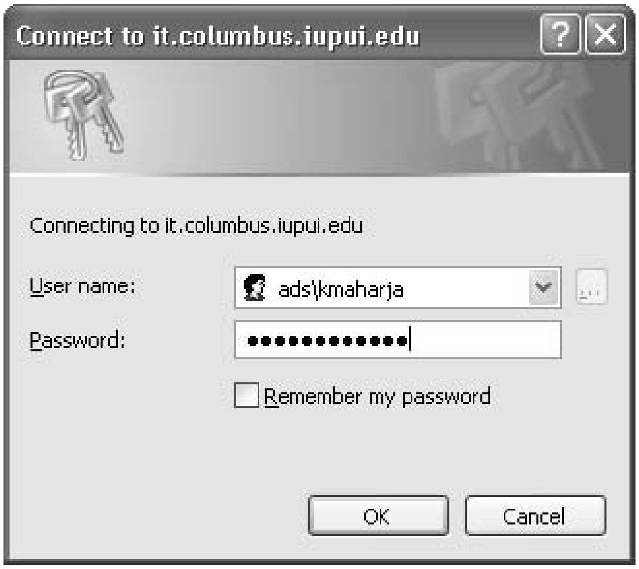
Downloading time depends on many factors. The following server, network, and users' machine characteristics affect downloading:

The server: Types of server software and hardware; server software configuration; number of people downloading from the same server at the same time.

The network: Noise in communication lines; traffic conditions on the Internet.

The user's machine: Load on the client machine (CPU + memory); load on the machine's network resources.

Obviously, the larger the size of a file, the more time it takes to complete the transfer. Therefore, it is better to verify the download time before files are actually downloaded. Some sites provide multipart copying, i.e., a large file available in a small number of smaller parts that can be downloaded and recombined on the user's machine. Figure 9 demonstrates transfer speed as 188 KB per second.



Connect to it.columbus.iupui.edu

Connecting to it.columbus.iupui.edu

User name: ads\kmaharja

Password: ••••••••••••

□ Remember my password

Figure 7: User name and password.

LIVE UPDATES AND SMART DOWNLOADS

"Live update" allows a running application to check automatically for updates via the Internet. Live updates automate the regular downloads and installation of software. If a personal computer is connected to the Internet and the computer is configured to check for the latest updates, the personal computer will automatically download and install the latest version of software, security patches, service packs, server releases, and other updates. Many companies such as Microsoft, Symantec, Caldera, RedHat, and RealNetworks provide live updates. Microsoft provides a Web-based Windows update which automatically scans for new updates, downloads necessary updates, and installs downloaded components with the user's permission. These updates include critical updates, security updates, Windows operating system specific updates, and driver updates. This Windows update Web site also lets individuals review and install updates of their choices via the Web. This is the most efficient way to keep users' computers current with the latest updates and upgrades. Some companies will just check if users have the latest version of the software installed. If an update is available, the users will have to download and install the software on their own. An example of live update is shown in Figure 10.

This example demonstrates an easy and effective way to keep a computer current with the latest version of antivirus software updates.

Many times, if the connection is broken during downloading or uploading, the connection needs to be reestablished and the downloading process reinitiated. Usually, this means the downloading process will start from the beginning again. This occurrence is especially frustrating if a large file was being downloaded via the slower connection speed. To circumvent this inconvenience, smart downloads can be implemented. Using a smart download, a download process will resume where it left off if the connection is broken during downloading. This is particularly advantageous for users who download via noisy, slower connections. Figure 11 demonstrates the usage of a smart download.

As can be seen from Figure 11, users can pause downloading and resume it later. This is particularly helpful when downloading large files over slow dial-up connections because, if for some reason the downloading were cancelled, downloading would resume from where it left off. Unfortunately, the user depends on the server to provide this feature, and it cannot be "invoked" by a user. Not all sites provide this functionality.

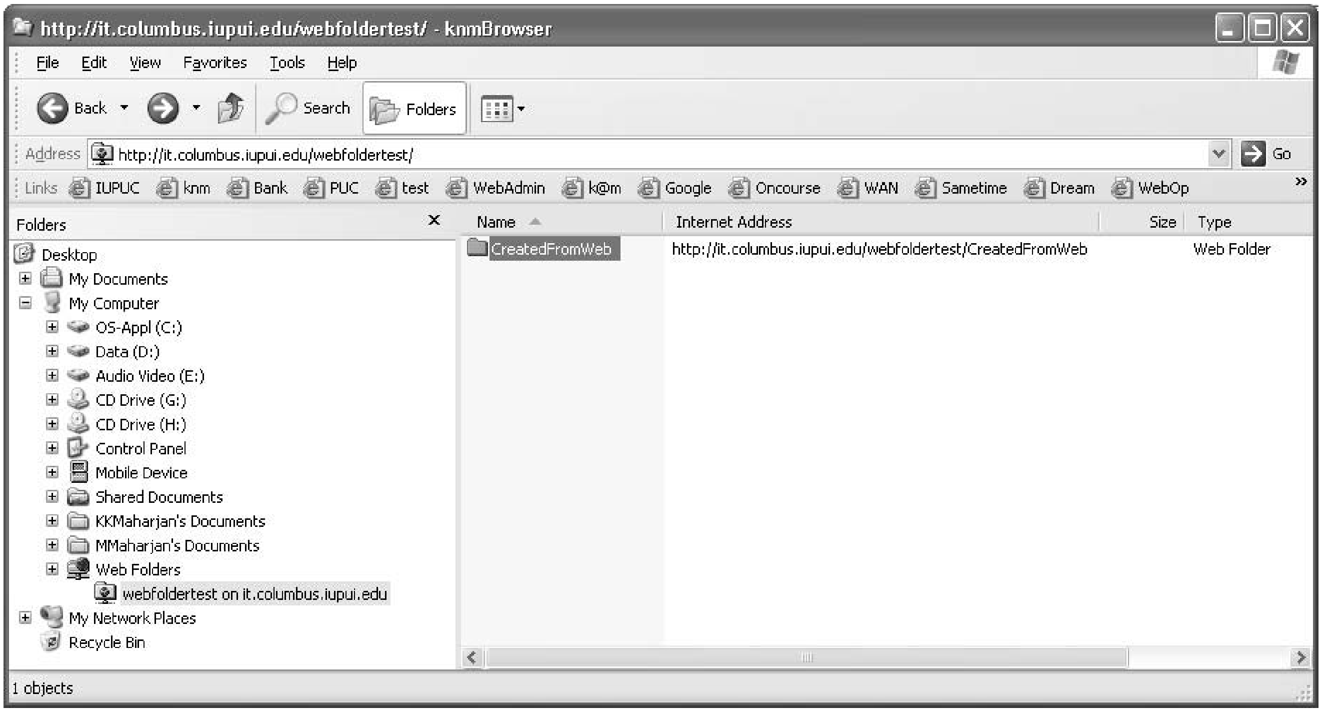
TYPICAL FILES FOR DOWNLOADING

Individuals can find several types of files to download from the Internet. The following list depicts some typical types of file available for downloading:

Audio files in WAV, MP3, WMP, and RA formats.

Image files in GIF, JPG, TIFF, and BMP formats.

Video files in AVI, RM, and MPG formats.



http://it.columbus.iupui.edu/webfoldertest/ - knmBrowser

Address http://it.columbus.iupui.edu/webfoldertest/

Name	Internet Address	Size	Type
CreatedFromWeb	http://it.columbus.iupui.edu/webfoldertest/CreatedFromWeb		Web Folder

1 objects

Figure 8: A Web site opened as a Web folder.

Computer-based training materials.

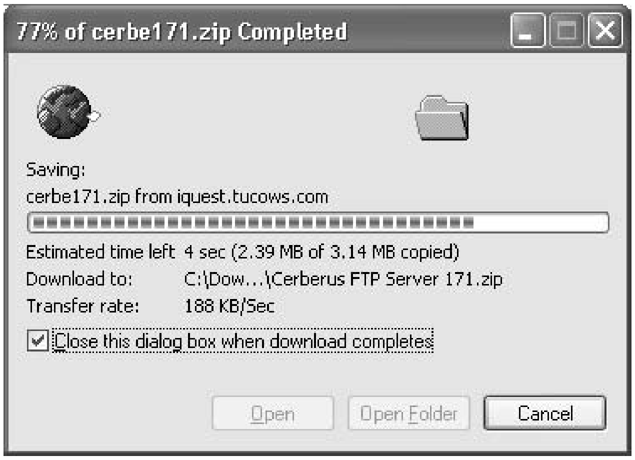
Documents and product information, manuals, tutorials, eBooks, eZines, maps, and directions in PDF, DOC, CSV, TXT, LIT, MPS, JPG, and HTM.

Various application programs in EXE and ZIP formats.

Various operating systems, e.g., Caldera Linux, RedHat Linux, and SuSe Linux in ISO formats.

Various service packs, service releases, security updates, critical updates, and application updates in different formats.

Streaming music, news, TV broadcasts, and lecture materials (both audio and video). Streaming data are not persistent on users' machines.



77% of cerbe171.zip Completed

Saving:
cerbe171.zip from iquest.tucows.com

Estimated time left 4 sec (2.39 MB of 3.14 MB copied)
Download to: C:\Dow...\Cerberus FTP Server 171.zip
Transfer rate: 188 KB/Sec

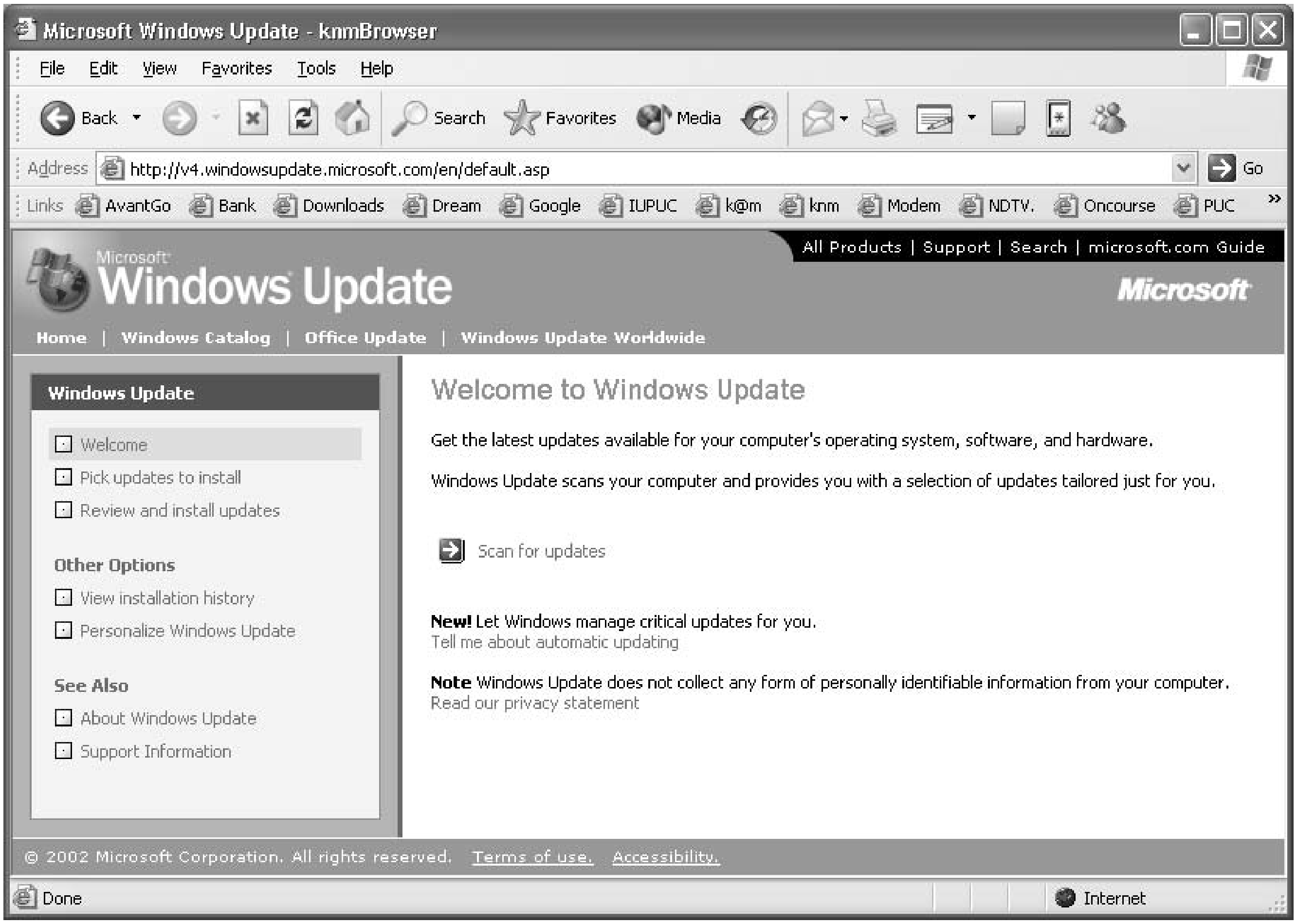
☑ Close this dialog box when download completes

Figure 9: Download speed (transfer rate).

Freeware/shareware/trialware/demoware software (e.g., applications, operating systems, games, utilities, device drivers, updates, and antivirus programs), flight information, stock quotes, and financial data are available for download. Users can download any type of file to any type of platform.

DOWNLOADING STEPS

This section describes the step-by-step procedure for downloading easily and efficiently. To organize the downloaded files properly, a folder should be created. This folder can be named anything—"Downloads," for example. As can be seen from Figure 4, subfolders can be created by company names, e.g., "Microsoft" and "Symantec." This enables easy access to the downloaded files in the future. Instead of naming subfolders by company name, subfolders can be organized by the type of the files as well as Operating Systems, Applications, Updates, Utilities, and AntiVirus. When connecting to an FTP server, users may be presented with a user name and a password request. Many FTP sites support anonymous FTP (a process of downloading files using file transfer protocol allowing the user to remain anonymous to the site from which the file is transferred) connections. This means users can designate the user name as "anonymous" and there is no password. Instead, as a courtesy, it is customary to send one's e-mail address as a password. If users are using a browser to connect to an FTP site, the browser handles an FTP download transparently for them. Anyone with an Internet connection can access files and folders at FTP sites that allow anonymous log-ins. This is an excellent way to disseminate information without maintaining any user names and passwords.


Microsoft Windows Update - knmBrowser
File Edit View Favorites Tools Help
Back Search Favorites Media
Address
Go
Links AvantGo Bank Downloads Dream Google IUPUC k@m knm Modem NDTV. Oncourse PUC
All Products | Support | Search | microsoft.com Guide
Microsoft Windows Update
Microsoft
Home | Windows Catalog | Office Update | Windows Update Worldwide
Windows Update
Welcome
Pick updates to install
Review and install updates
Other Options
View installation history
Personalize Windows Update
See Also
About Windows Update
Support Information
Scan for updates
Done
Internet

http://v4.windowsupdate.microsoft.com/en/default.asp

Welcome to Windows Update

Get the latest updates available for your computer's operating system, software, and hardware.

Windows Update scans your computer and provides you with a selection of updates tailored just for you.

New! Let Windows manage critical updates for you.
Tell me about automatic updating

Note Windows Update does not collect any form of personally identifiable information from your computer.
Read our privacy statement

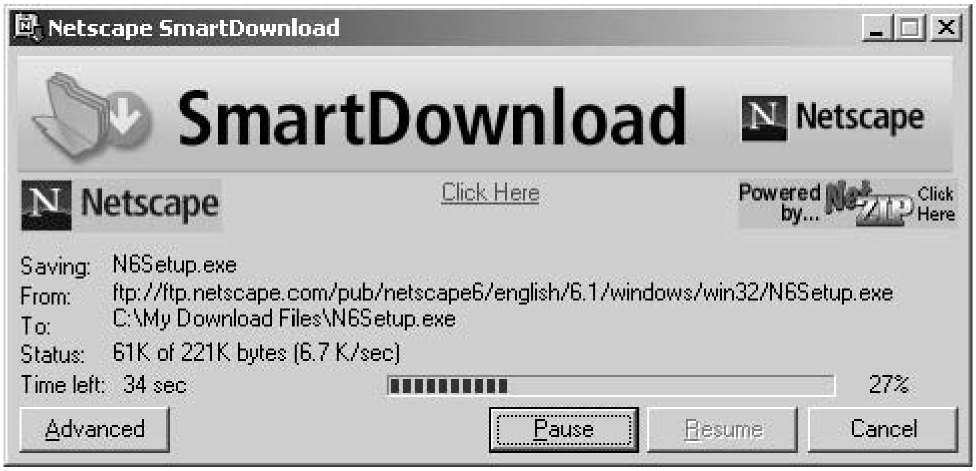
© 2002 Microsoft Corporation. All rights reserved. Terms of use. Accessibility.

Figure 10: Microsoft version of live update called "Windows update."

Creating a Download Folder

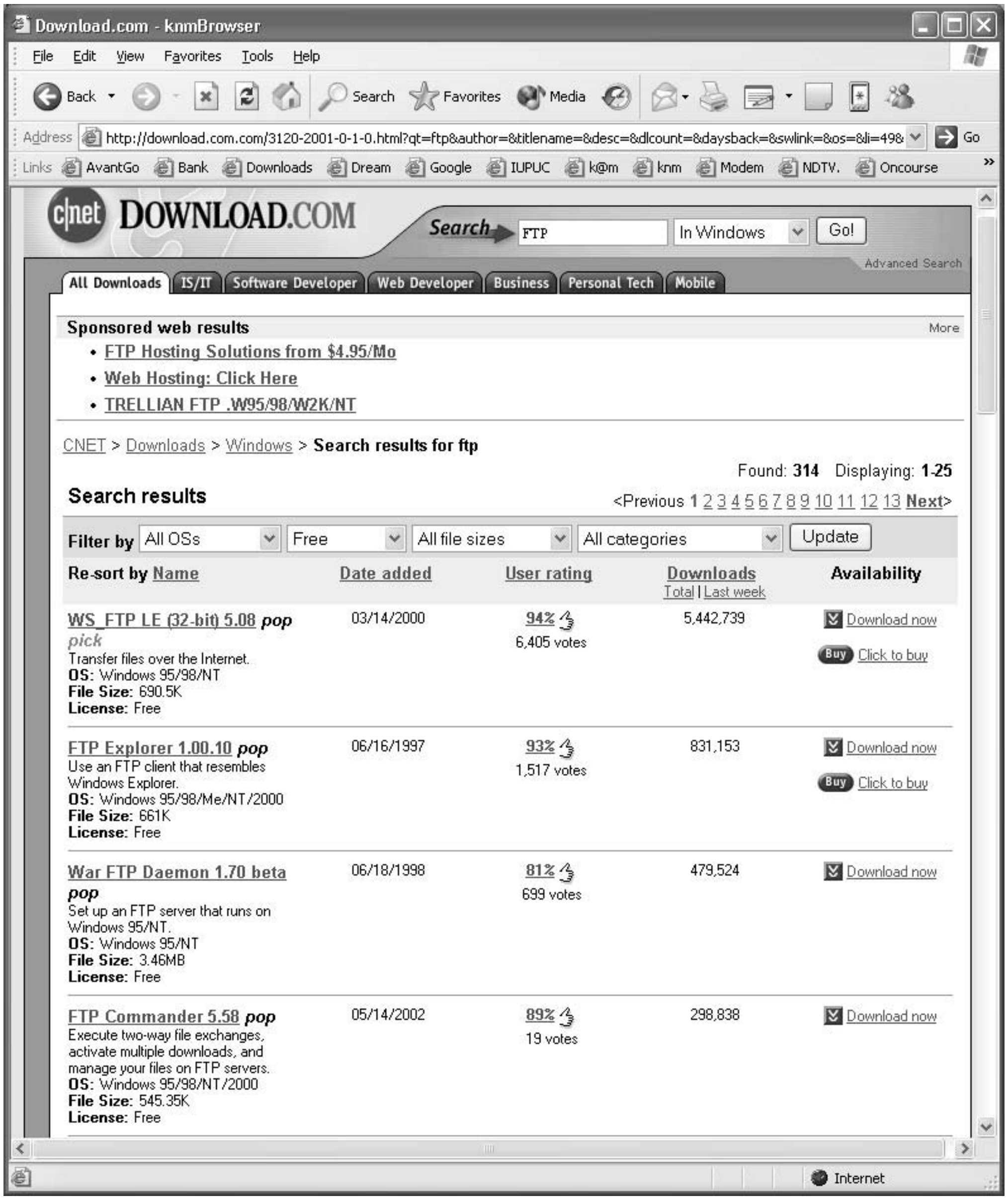
It is best to create a separate folder to store all the downloaded files. Depending upon the operating systems, the process of creating a directory may vary. In a Windows environment, a new directory can be created by using the following steps:

(1) Right click on "Start."
(2) Click "Explore" to open the Windows Explorer.
(3) Click on the drive letter (e.g., C:) to select the drive where the folder to receive downloaded files is to go.
(4) Right click on any blank space on the Windows explorer—a pop-up menu will open.



Saving: N6Setup.exe
From: ftp://ftp.netscape.com/pub/netscape6/english/6.1/windows/win32/N6Setup.exe
To: C:\My Download Files\N6Setup.exe
Status: 61K of 221K bytes (6.7 K/sec)
Time left: 34 sec

Figure 11: Using a smart download.


Download.com - knmBrowser
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://download.com.com/3120-2001-0-1-0.html?qt=ftp&author=&titlename=&desc=&dlcount=&daysback=&swlink=&os=&li=49& Go
Links AvantGo Bank Downloads Dream Google IUPUC k@m knm Modem NDTV. Oncourse
cnet DOWNLOAD.COM
Search FTP In Windows Go!
Advanced Search
All Downloads IS/IT Software Developer Web Developer Business Personal Tech Mobile
Internet

Sponsored web results More

- FTP Hosting Solutions from $4.95/Mo
- Web Hosting: Click Here
- TRELLIAN FTP .W95/98/W2K/NT

CNET > Downloads > Windows > Search results for ftp

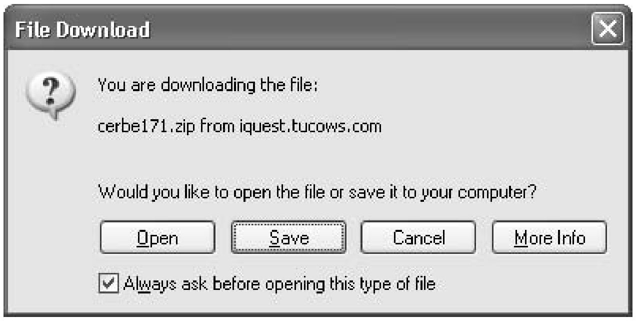
Found: 314 Displaying: 1-25

Search results <Previous 1 2 3 4 5 6 7 8 9 10 11 12 13 Next>

Filter by All OSs Free All file sizes All categories Update

Re-sort by Name	Date added	User rating	Downloads Total | Last week	Availability
WS_FTP LE (32-bit) 5.08 *pop pick* Transfer files over the Internet. **OS:** Windows 95/98/NT **File Size:** 690.5K **License:** Free	03/14/2000	94% 6,405 votes	5,442,739	Download now Buy Click to buy
FTP Explorer 1.00.10 *pop* Use an FTP client that resembles Windows Explorer. **OS:** Windows 95/98/Me/NT/2000 **File Size:** 661K **License:** Free	06/16/1997	93% 1,517 votes	831,153	Download now Buy Click to buy
War FTP Daemon 1.70 beta *pop* Set up an FTP server that runs on Windows 95/NT. **OS:** Windows 95/NT **File Size:** 3.46MB **License:** Free	06/18/1998	81% 699 votes	479,524	Download now
FTP Commander 5.58 *pop* Execute two-way file exchanges, activate multiple downloads, and manage your files on FTP servers. **OS:** Windows 95/98/NT/2000 **File Size:** 545.35K **License:** Free	05/14/2002	89% 19 votes	298,838	Download now

Figure 12: Browsing the Internet for the file.



You are downloading the file:

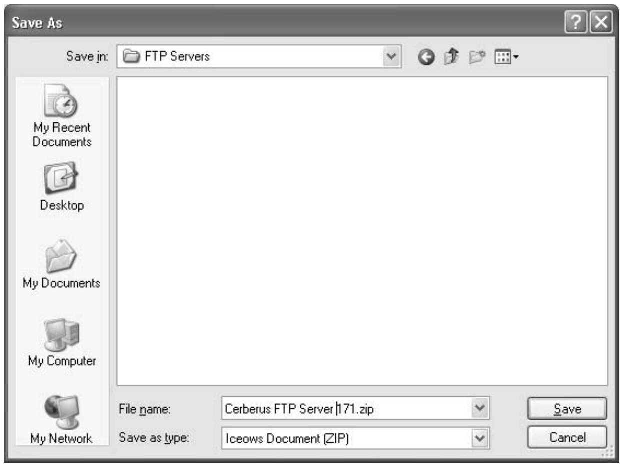
cerbe171.zip from iquest.tucows.com

Would you like to open the file or save it to your computer?

Open Save Cancel More Info

☑ Always ask before opening this type of file

Figure 13: File download dialog box.

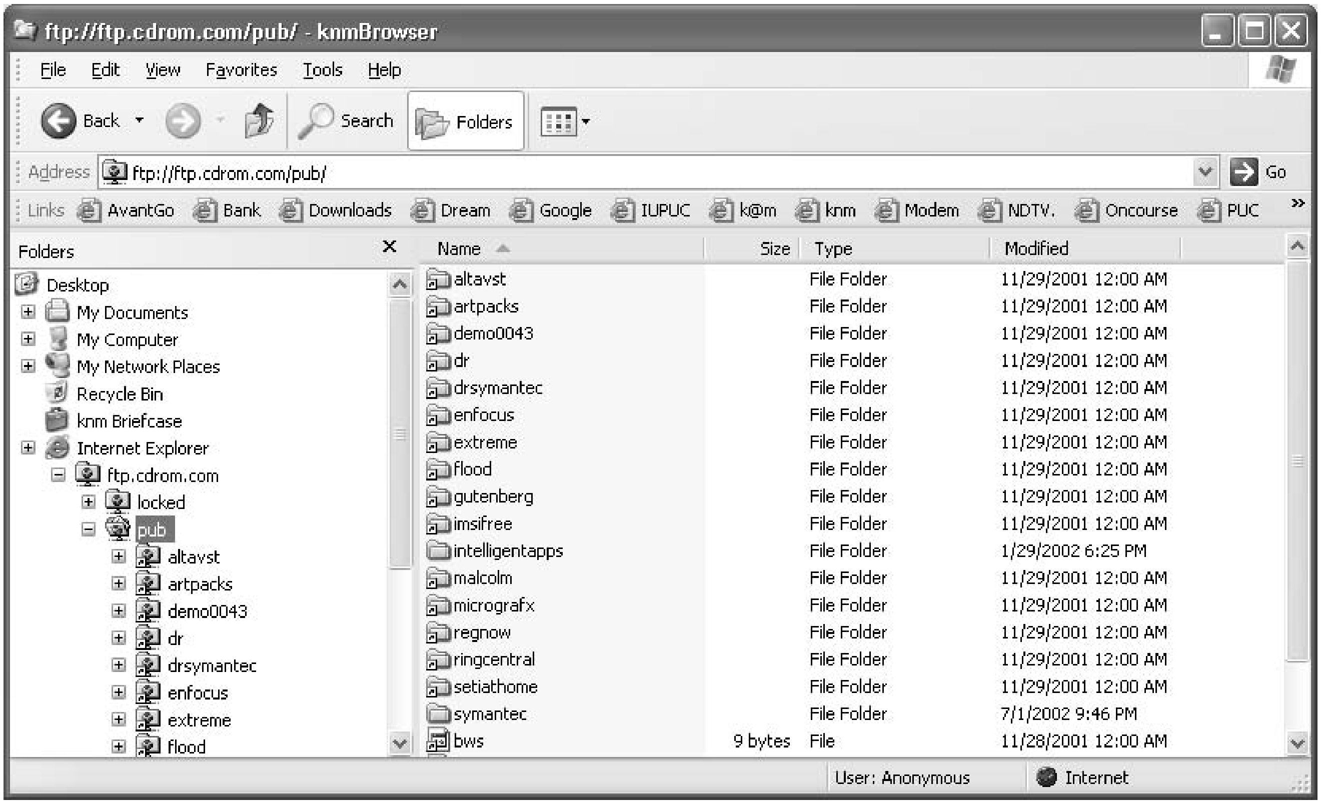


Save in: FTP Servers

File name: Cerberus FTP Server171.zip

Save as type: Iceows Document (ZIP)

Figure 14: Renaming and saving a file.



Address: ftp://ftp.cdrom.com/pub/

Name	Size	Type	Modified
altavst		File Folder	11/29/2001 12:00 AM
artpacks		File Folder	11/29/2001 12:00 AM
demo0043		File Folder	11/29/2001 12:00 AM
dr		File Folder	11/29/2001 12:00 AM
drsymantec		File Folder	11/29/2001 12:00 AM
enfocus		File Folder	11/29/2001 12:00 AM
extreme		File Folder	11/29/2001 12:00 AM
flood		File Folder	11/29/2001 12:00 AM
gutenberg		File Folder	11/29/2001 12:00 AM
imsifree		File Folder	11/29/2001 12:00 AM
intelligentapps		File Folder	1/29/2002 6:25 PM
malcolm		File Folder	11/29/2001 12:00 AM
micrografx		File Folder	11/29/2001 12:00 AM
regnow		File Folder	11/29/2001 12:00 AM
ringcentral		File Folder	11/29/2001 12:00 AM
setiathome		File Folder	11/29/2001 12:00 AM
symantec		File Folder	7/1/2002 9:46 PM
bws	9 bytes	File	11/28/2001 12:00 AM

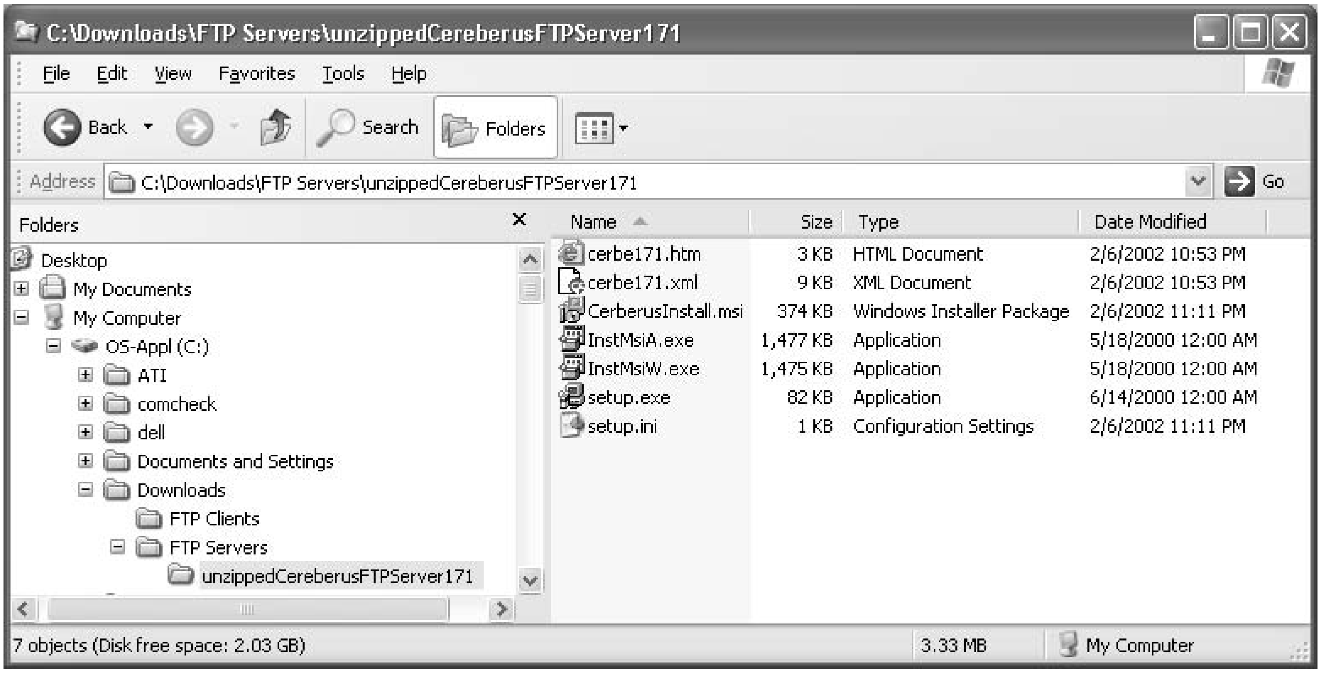
Figure 15: Downloading files and folders using FTP and a Web browser.

(5) Click on “New” on the pop-up menu to create a folder—the operating system will create a new folder and give it a name of New Folder.

(6) Rename the folder to “Downloads.”

(7) Click at any blank area to complete the task.

(8) Repeat steps 1 through 7 to create new subfolders. On step 3, instead of clicking on the drive letter, click on the “Downloads” folder that was just created.



Address: C:\Downloads\FTP Servers\unzippedCereberusFTPServer171

Name	Size	Type	Date Modified
cerbe171.htm	3 KB	HTML Document	2/6/2002 10:53 PM
cerbe171.xml	9 KB	XML Document	2/6/2002 10:53 PM
CerberusInstall.msi	374 KB	Windows Installer Package	2/6/2002 11:11 PM
InstMsiA.exe	1,477 KB	Application	5/18/2000 12:00 AM
InstMsiW.exe	1,475 KB	Application	5/18/2000 12:00 AM
setup.exe	82 KB	Application	6/14/2000 12:00 AM
setup.ini	1 KB	Configuration Settings	2/6/2002 11:11 PM

7 objects (Disk free space: 2.03 GB) 3.33 MB

Figure 16: Uncompressing the downloaded file.

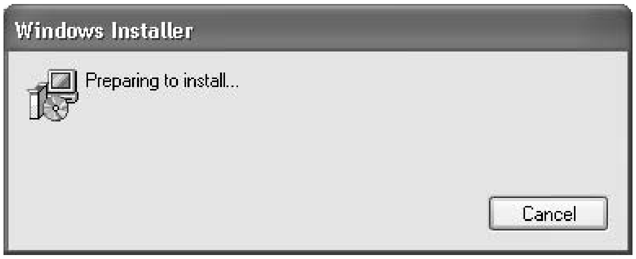
Browsing the Internet for the File

Once the directory structure has been created for receiving files and folders downloaded from the Web, it is time to browse the Internet for the files to be downloaded. Depending upon the types of files needed to be downloaded, one can start browsing Web sites such as http://download.com, http://www.webmasterfree.com, http://www.tucows.com, http://yahoo.com, or http://google.com. If users discover something that looks interesting and they decide to download, an appropriate directory structure can be created at this time as well. These sites are very well organized by category of software, e.g., operating systems, MP3 and audio, Internet, games, business, mobile, multimedia & design, Web developer, software developer, utilities & drivers, and home & desktop. Users should search for the files in the desired category. If the user is looking for a particular file from a specific company, then the best action would be to visit the company's home page and look for the necessary file. Search engines such as http://www.google.com and http://www.yahoo.com assist in finding the companies as well. Using this site, users can search for files by operating system, license type (free and free to try), size, and category. For an example, if a user is looking for freeware only and has selected "free" instead of the default "all licenses," then the http://download.com Web site will automatically filter out other types of files. Figure 12 shows the search result of a search "FTP for Windows" at http://download.com: 314 related programs, all free, are available for download.

Downloading a File Using a Browser or FTP Client Software

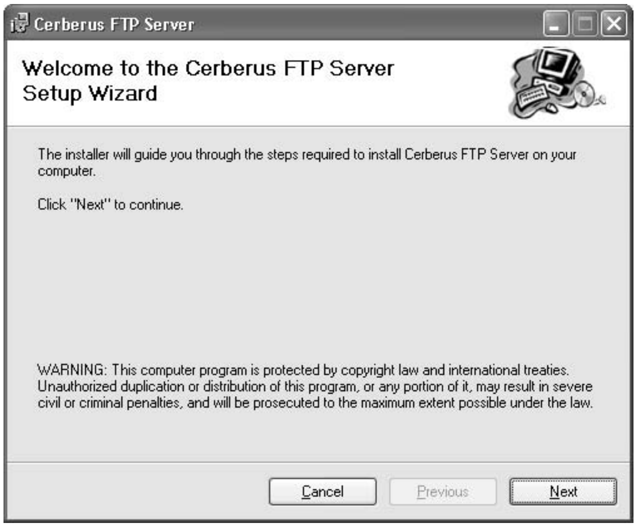
Once the desired file has been located, it can be downloaded using a browser or FTP client software. Most browsers support FTP in addition to HTTP. Downloading using a browser is quite simple. One should simply click on the desired file for downloading. A "file download" dialog box will be opened (Figure 13).

Even though users can save a file without renaming it, it is preferable to rename a file being downloaded. For example, in Figures 13 and 14 the original file cerbe171.zip is renamed to Cerberus FTP Server 171.zip and the file is being saved in the C:\Downloads\FTP Servers\ directory. Figure 9 depicts the actual transfer of the file. After some time the user may forget what cerbe171.zip is. However, Cerberus FTP Server 171.zip will provide all the information the user needs. In this case, this is FTP Server software version 1.71 from Cerberus. This way the user will always be able to find the necessary files quickly and efficiently. Once the application has been installed, users no longer need the downloaded file; however, it is wise to keep a particular release if the users need to revert to a previous version should a new one prove to be unstable.


Windows Installer

Preparing to install...

Figure 17: Windows Installer preparing to install.


Cerberus FTP Server

Welcome to the Cerberus FTP Server Setup Wizard

The installer will guide you through the steps required to install Cerberus FTP Server on your computer.

Click "Next" to continue.

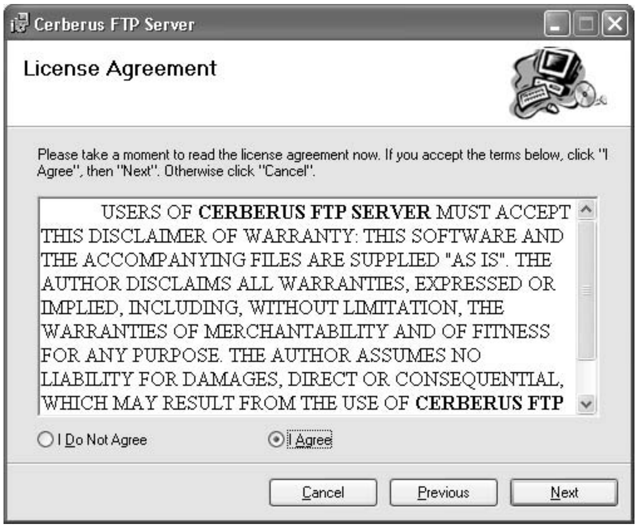
WARNING: This computer program is protected by copyright law and international treaties. Unauthorized duplication or distribution of this program, or any portion of it, may result in severe civil or criminal penalties, and will be prosecuted to the maximum extent possible under the law.

Figure 18: Setup Wizard.

Figure 15 illustrates the files and folders available for downloading. This interface is very similar to opening a Web site as a folder. To download files and folders, all the users have to do is drag and drop. The prefix "ftp://" in ftp://ftp.cdrom.com/.1/gutenberg/images/ at the address bar in Figure 15 tells the browser that the FTP protocol is being used. The browser has connected to the ftp.cdrom.com FTP server using "anonymous" as the log-in name transparently.

Preparing for Installation

Once necessary files have been downloaded, it is time to install the programs. If the downloaded file is a non-self-extracting compressed file, users should uncompress


Cerberus FTP Server

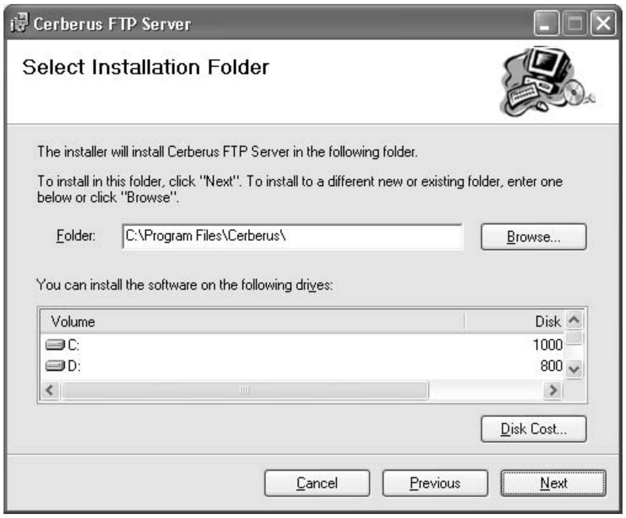
License Agreement

Please take a moment to read the license agreement now. If you accept the terms below, click "I Agree", then "Next". Otherwise click "Cancel".

USERS OF **CERBERUS FTP SERVER** MUST ACCEPT THIS DISCLAIMER OF WARRANTY: THIS SOFTWARE AND THE ACCOMPANYING FILES ARE SUPPLIED "AS IS". THE AUTHOR DISCLAIMS ALL WARRANTIES, EXPRESSED OR IMPLIED, INCLUDING, WITHOUT LIMITATION, THE WARRANTIES OF MERCHANTABILITY AND OF FITNESS FOR ANY PURPOSE. THE AUTHOR ASSUMES NO LIABILITY FOR DAMAGES, DIRECT OR CONSEQUENTIAL, WHICH MAY RESULT FROM THE USE OF **CERBERUS FTP**

○ I Do Not Agree ● I Agree

Figure 19: Agreeing to the License Agreement.



Select Installation Folder

The installer will install Cerberus FTP Server in the following folder.

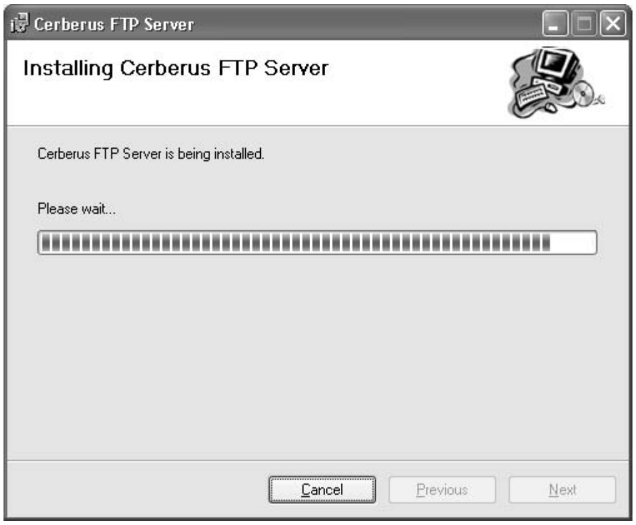
To install in this folder, click "Next". To install to a different new or existing folder, enter one below or click "Browse".

Folder: C:\Program Files\Cerberus\

You can install the software on the following drives:

Volume	Disk
C:	1000
D:	800

Figure 20: Selecting the installation folder. The default drive and folders are being selected.



Installing Cerberus FTP Server

Cerberus FTP Server is being installed.

Please wait...

Figure 22: Actual installation process.

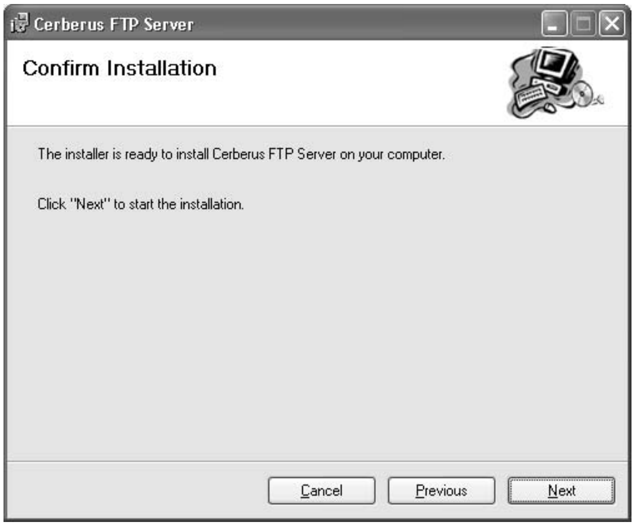
(unzip) it first. The first task is to decide where to install the program, how much space is necessary to install the downloaded software, and how much space is available. After this, the installation process is straightforward. Many times unzipping a file will automatically invoke the installation procedure, whereas in some cases, users will have to go to the folder into which the unzipped (uncompressed) files were placed and run a special program usually called setup.exe. Figure 16 depicts all uncompressed files from the Cerberus FTP Server 171.zip file in a subfolder called unzippedCereberusFTPServer171.

Opening Downloaded File

It is best to organize downloaded files as neatly as possible. This definitely takes some extra effort in the beginning, but it will save you a lot of time and anxiety later. Think of two rooms. In the first room all personal belongings are spread (stored) everywhere on the floor. The second room is very well organized with all personal belongings stored neatly by category. It is much easier to store items in the first room because all one has to do is throw one's belongings somewhere on the floor whereas in the second the room one has to think where to store before storing properly. When it is time to find one's belongings, it is very easy to find them in the second room because one knows exactly where the personal belongings are stored. The same principle applies in storing downloaded files.

Installing Downloaded File

In the final step, double click on the setup.exe file in the subfolder C:\Downloads\FTP Servers\unzipped CereberusFTPServer171. The actual installation procedures may vary from program to program but usually the installer will guide individuals through the whole installation procedure by means of the steps shown in Figures 17–24.

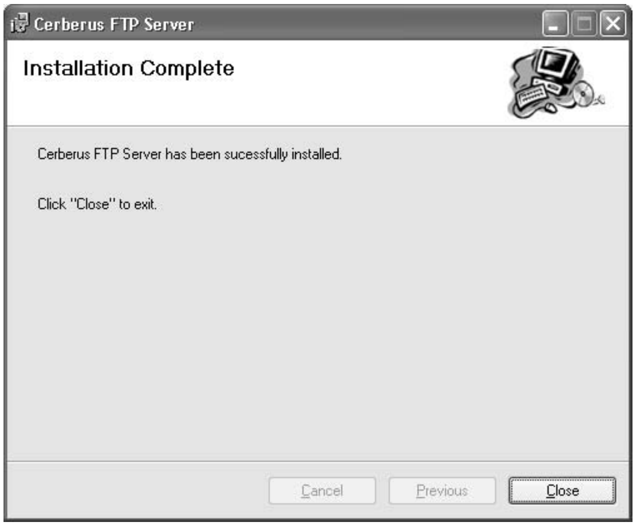


Confirm Installation

The installer is ready to install Cerberus FTP Server on your computer.

Click "Next" to start the installation.

Figure 21: Confirming the installation.



Installation Complete

Cerberus FTP Server has been sucessfully installed.

Click "Close" to exit.

Figure 23: Installation complete. Click on "Close"

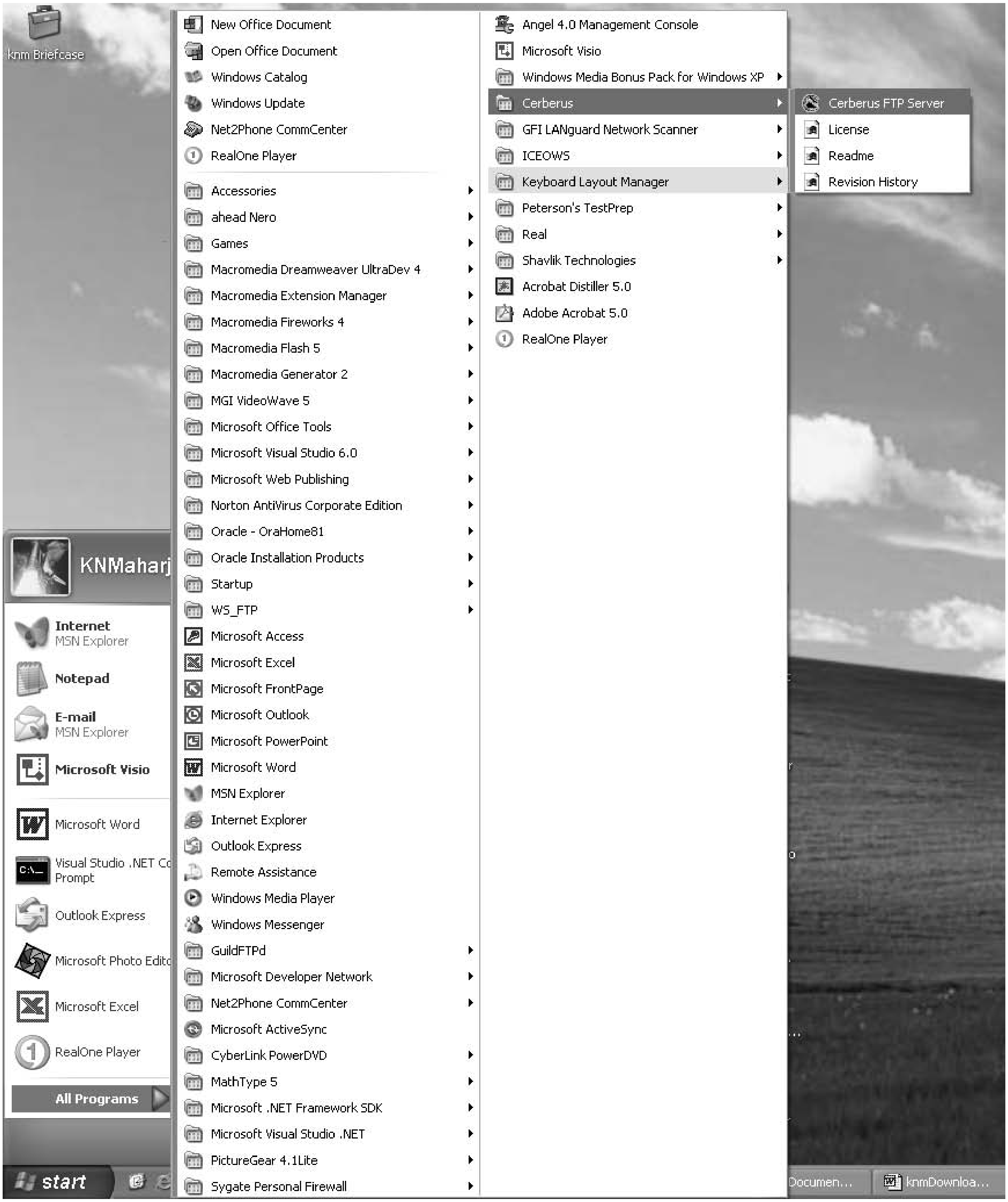
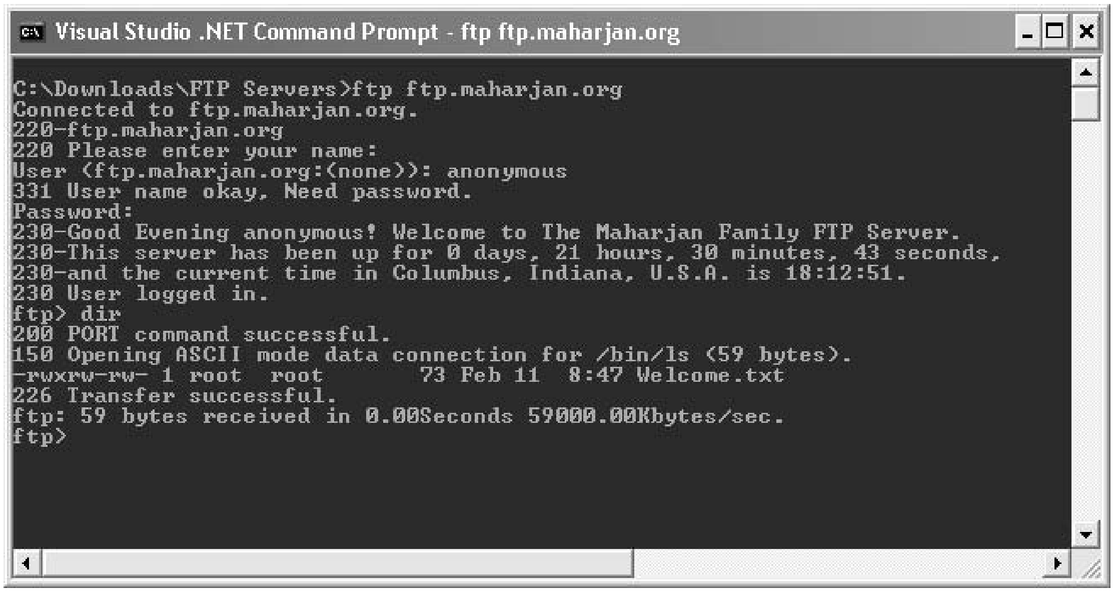


Figure 24: Invoking the program after successful installation.



```
C:\Downloads\FTP Servers>ftp ftp.maharjan.org
Connected to ftp.maharjan.org.
220-ftp.maharjan.org
220 Please enter your name:
User (ftp.maharjan.org:(none)): anonymous
331 User name okay, Need password.
Password:
230-Good Evening anonymous! Welcome to The Maharjan Family FTP Server.
230-This server has been up for 0 days, 21 hours, 30 minutes, 43 seconds,
230-and the current time in Columbus, Indiana, U.S.A. is 18:12:51.
230 User logged in.
ftp> dir
200 PORT command successful.
150 Opening ASCII mode data connection for /bin/ls (59 bytes).
-rwxrw-rw- 1 root  root       73 Feb 11  8:47 Welcome.txt
226 Transfer successful.
ftp: 59 bytes received in 0.00Seconds 59000.00Kbytes/sec.
ftp>
```

Figure 25: Character-based FTP client software.

To run the recently installed program (in this example), simply click on "Start," select "Cerberus," and then select "Cerberus FTP Server." If the installer has created a shortcut on the desktop, the user can double click on the shortcut to invoke the program as well.

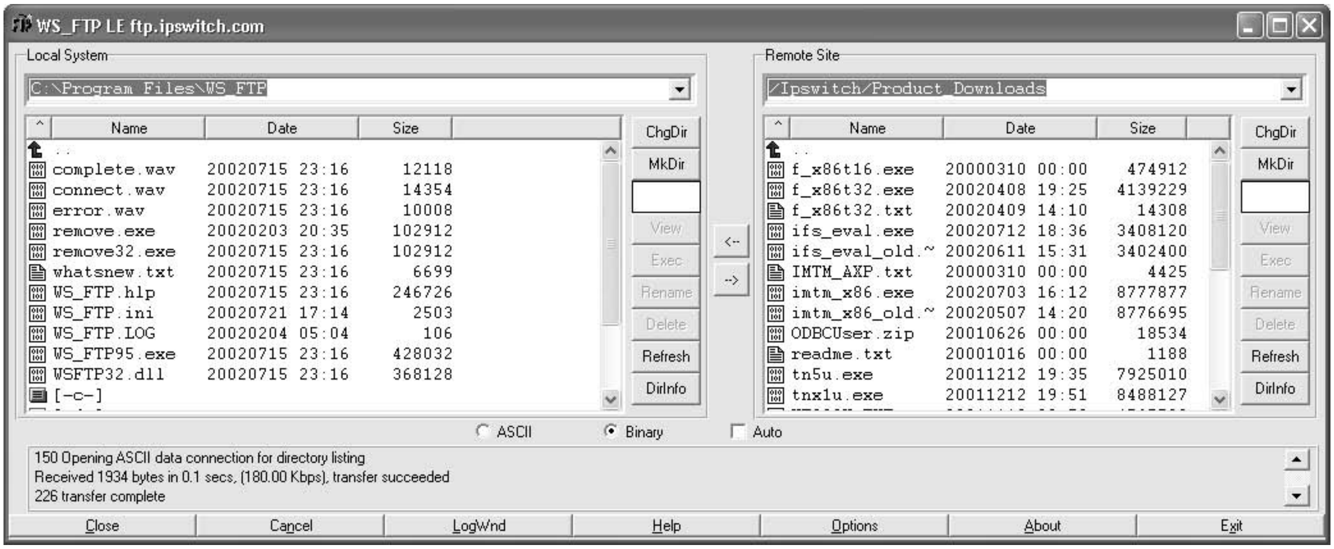
VARIOUS FTP CLIENT SOFTWARE PROGRAMS

If users choose to use an FTP client explicitly, rather than a Web browser, an alternative way to download files and folders is by using FTP. Before GUI-based FTP clients came over the horizon, users had to use character-based FTP client software to download files (Figure 25). For most users, this software is not user-friendly because users need to remember FTP commands to interact with it. Character-based FTP client software is distributed with today's Windows and other operating systems.

Fortunately, a number of good FTP client programs are available today, both freely and commercially. Virtually all of these software products are GUI-based (see, e.g., Figure 26). Compared to character-based FTP client software, these GUI-based programs are accessible, efficient, and intuitive. Some examples are

Freeware FTP programs: WS_FTP LE, GoZilla Free, SmartFTP, FTP Explorer, LeachFTP, Net Vampire, FTP Commander, Fictional Daemon, ES Download Resumer, HS FTPExplorer, and DLExpert.

Shareware FTP programs: Bullet Proof FTP, Internet Neighborhood, ftpNetDrive, 1stChoice FTPPro,



```
C:\Program Files\WS_FTP
Name              Date              Size
..
complete.wav      20020715 23:16    12118
connect.wav       20020715 23:16    14354
error.wav         20020715 23:16    10008
remove.exe        20020203 20:35   102912
remove32.exe      20020715 23:16   102912
whatsnew.txt      20020715 23:16     6699
WS_FTP.hlp        20020715 23:16   246726
WS_FTP.ini        20020721 17:14     2503
WS_FTP.LOG        20020204 05:04      106
WS_FTP95.exe      20020715 23:16   428032
WSFTP32.dll       20020715 23:16   368128
[-c-]
```

```
/Ipswitch/Product_Downloads
Name              Date              Size
..
f_x86t16.exe      20000310 00:00   474912
f_x86t32.exe      20020408 19:25  4139229
f_x86t32.txt      20020409 14:10    14308
ifs_eval.exe      20020712 18:36  3408120
ifs_eval_old.~    20020611 15:31  3402400
IMTM_AXP.txt      20000310 00:00     4425
imtm_x86.exe      20020703 16:12  8777877
imtm_x86_old.~    20020507 14:20  8776695
ODBCUser.zip      20010626 00:00    18534
readme.txt        20001016 00:00     1188
tn5u.exe          20011212 19:35  7925010
tnx1u.exe         20011212 19:51  8488127
```

```
150 Opening ASCII data connection for directory listing
Received 1934 bytes in 0.1 secs, (180.00 Kbps), transfer succeeded
226 transfer complete
```

Figure 26: WS_FTP LE, GUI-based FTP client software.

BlackWidow, FlashFXP, LeapFTP, FTP Now, 3D-FTP, and AceFTP.

Public domain programs: llnlxftp, moxftp, and ncftp.

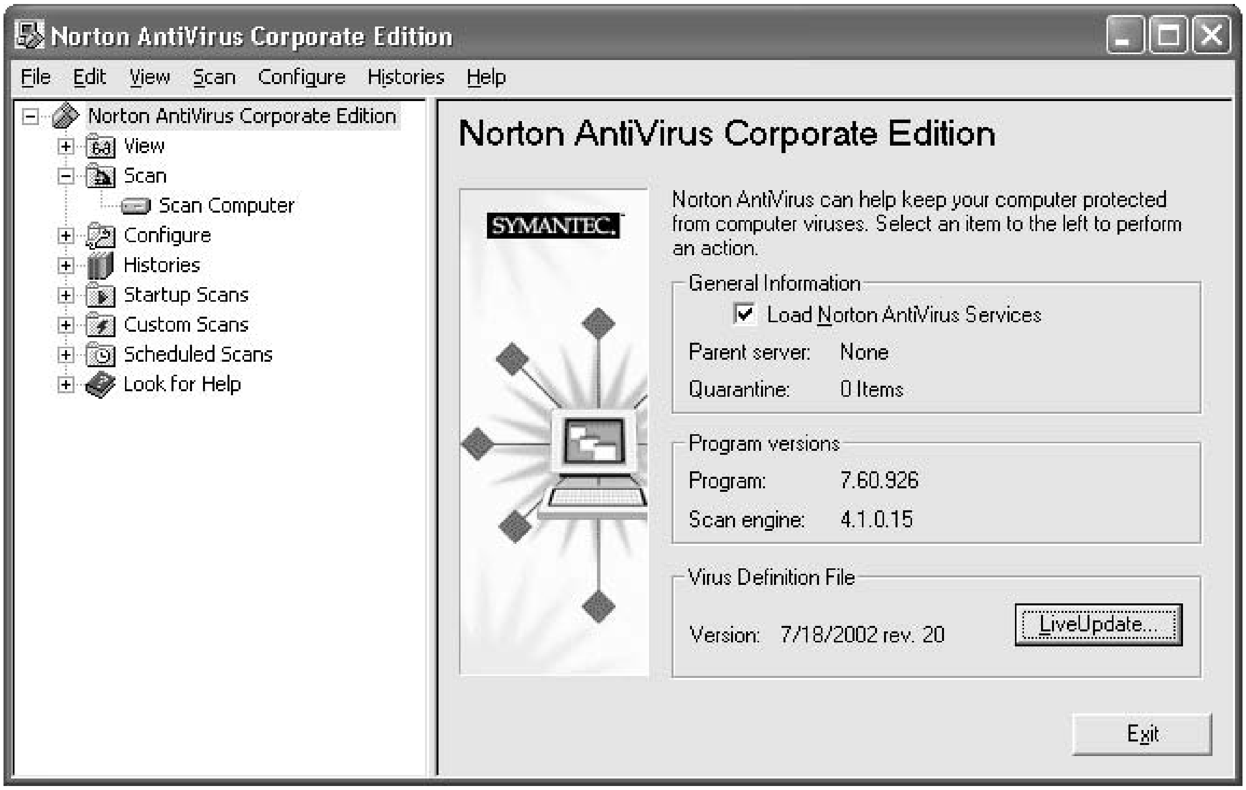
Among these programs WS_FTP LE seems to be very popular because this program is provided free of charge for a home (nonbusiness) user; a federal, state, or local government employee in the United States; and a student, faculty, or staff member at an educational institution (K-12, junior college, college, or university).

SAFETY ISSUES OF DOWNLOADING

Freely available programs are not necessarily legal or safe to download. The main concern in downloading from the Internet is the possibility of getting computer viruses via the Web. A hidden spyware/adware program may be bundled in freeware and shareware. (This type of software secretly collects user information through the user's Internet connection for advertising purposes without the user's knowledge. Usually, this software is bundled as a hidden portion of freeware or shareware and installed automatically without the users' knowledge.) Even the programs downloaded directly from popular sites may contain Trojan-horse viruses. For instance, recently:

> Two popular file-swapping sites have been unknowingly spreading a Trojan horse that sends user information such as IDs and IP addresses to a third party.
>
> The Trojan, called "W32.DIDer," is installed on a user's computer as part of the normal installation process for Grokster and LimeWire—two freeware apps that emulate Napster's file-swapping capabilities. . . .
>
> Many freeware apps have promotional software bundled with them—it helps pay the bills—but usually the promotions are simply used to flash up advertising or lure users to a website. But this promotion, called "Clicktilluwin," secretly installs a program that carries the Trojan. During the installation processes of these freeware programs, you are asked if you want to install Clicktilluwin. Regardless of whether you click yes or no, the Trojan code is installed. The code makes changes to your PC's registry so the Trojan runs each time you start Windows. The Trojan then sends information like your user ID, your IP address, and your browser preferences to a website, 2001–07.com, which has since been shut down by the ISP.
>
> Both LimeWire and Grokster have issued warnings and apologies about the Trojan. LimeWire has released a new version of its software without the Trojan.
>
> Grokster has a patch on its site at http://www.techtv.com/news/security/story/0,24195,3366715,00.html. (Worley, 2002)

These malicious viruses may be programmed to do anything from wiping out hard drives to sending information from a computer to an attacker's computer or gaining access to other computers via the compromised



Norton AntiVirus Corporate Edition

Norton AntiVirus can help keep your computer protected from computer viruses. Select an item to the left to perform an action.

General Information

☑ Load Norton AntiVirus Services

Parent server: None

Quarantine: 0 Items

Program versions

Program: 7.60.926

Scan engine: 4.1.0.15

Virus Definition File

Version: 7/18/2002 rev. 20

Figure 27: Norton AntiVirus LiveUpdate.

computer. Therefore, users must be particularly careful before downloading from the Internet.

How can the user prevent a computer from being infected? One way to prevent this is by not opening files directly by applications. For example, if the user needs to open an attached Microsoft Word file, the user first downloads the file and then runs an antivirus program to check for a possible virus on every downloaded file after downloading. Only then should the user open the downloaded document.

It is also very important that the antivirus software itself is kept current. Most antivirus software companies provide automated updates of their software and virus definitions. All the user needs to do is to make sure that he/she has scheduled a live update and the computer is connected to the Internet to keep the antivirus software updated at the scheduled day and time. Figure 27 depicts the Norton AntiVirus LiveUpdate. This antivirus software can be scheduled to check for updated programs and download updates, and install updated programs and virus definition files automatically without any user intervention.

LEGAL ASPECTS OF DOWNLOADING

When researching the Internet, users will probably find a variety of commercial software, music, movies, pictures (images), personal information, as well as access to other commercial resources and even CD Keys and Key Generator programs for specific software. These programs are usually provided by hackers and code (software) crackers. Are these legal? Downloading illegally provided commercial software is definitely not legal. The sites would be closed quickly. Were there not so many new sites emerging, it would be easier to detect and dismantle all illegal sites immediately. The users learning the fundamentals of downloading from the Web must also comprehend the ethics of downloading. Software companies around the world are losing billions of dollars every year due to software piracy.

CONCLUSION

With the advent of ubiquitous, cheap dial-up connections, ever expanding broadband connections, worldwide networking of virtually every computer, and the development of GUI-based intuitive and user-friendly software programs, downloading files and folders from the Internet has become easy and popular. Because of the simplicity of locating files on the Internet, using search engines, and the ability to download files with a few clicks, more and more users are downloading a variety of files every day. Millions of servers worldwide are providing all kinds of audio, video, image, and text files from legal to illegal and from clean to virus loaded. Therefore, before downloading files to their computers, users should really know what the implications to their computers or contents of their computers might be if virus-loaded software were downloaded. While it is very important to organize downloaded files and to scan for a virus on downloaded files, it is equally important to make sure that users are aware of the legal aspects of downloading as well.

GLOSSARY

File Extensions

.avi Audio video interleaved animation file (video for Windows).

.bmp Bit-mapped, the file format for graphics in Windows.

.csv Comma separated value format. ASCII (American standard code for information interchange. This code is used for representing English characters.) text file format with fields separated by a comma.

.doc A document created by Microsoft Word word processor. Using Internet Explorer, these documents can be viewed directly in the Web browser provided the Microsoft Word program is installed.

.exe An executable file (a computer program).

.gif Graphic interchange format, a still and an animated image file format for storing images.

.htm/html A document conforming to hyper text markup language format.

.jpg or .jpeg Joint photographic experts group, an image file format developed by Joint Motion Picture Group and extensively used in the Internet.

.mov A QuickTime video clip.

.mp3 MPEG audio layer 3 used to compress audio signals.

.mpg Video format conforming to motion picture expert group (MPEG) standard.

.png Portable network graphics, a patent-free graphics standard.

.pdf Portable document format, a popular document distribution format developed by Adobe. A specialized program such as Adobe Acrobat reader is required to read or print these types of documents.

.ra Real audio format developed by Real Networks to compress audio and music. This format supports streaming technology to deliver music and news via the Internet.

.rm Real movie format developed by Real Networks to compress video.

.sit Compressed Macintosh archive created by STUFFIT.

.tif A tagged images file format (TIFF) file, a popular file format for storing bit-mapped images on PCs and Macintosh.

.txt A file containing only ASCII text.

.wav Waveform audio file (RIFF WAVE format) for Windows.

.wma Windows Media Audio developed by Microsoft. Microsoft boasts that it is capable of compressing an MP3 file up to one-third of its size without diminishing audio quality. Using this technology up to 300 individual songs can be recorded in a regular recordable CD-ROM.

.zip A compressed file created using utilities such as WinZip, PKZip, and UltimateZip.

Terms

Anonymous FTP A process of downloading files using file transfer protocol (FTP) that allows the user to remain anonymous to the site from which the file is transferred.

ASCII American standard code for information interchange. This code is used for representing English characters.

Baud The number of signaling events per second. At lower speeds, only one bit of information (signaling event) is encoded in each electrical change. However, at higher speeds, it is possible to encode more than one bit in each electrical change.

Bandwidth A connection speed between a client machine and a server machine, usually measured in Kbps or Mbps.

Bit The smallest unit of information either 0 or 1. It stands for BInary digT.

Browser Short for Web browser, a software application used to find and display Web pages.

Byte Usually eight bits of information grouped together. This is the basic unit of data transferred during a download and is also the data unit containing one ASCII character.

Computer-based training A student learns by running special educational programs on a computer.

Client A computer running a specialized client component of client/server software to request/receive services from a server.

Cookies Information provided to a Web browser by a Web server and stored on a client computer in a text file. Next time the user visits the same Web server this information will be sent to the Web server. The main purpose of cookies is to identify users and their tastes to customize Web pages for them.

Commercial software Software which is available for a fee.

Freeware Copyrighted software provided for free. Usually, even though this software is provided for free it cannot be modified or sold.

FTP File transfer protocol developed to transfer files among networked computers.

HTML Hypertext markup language used to create documents on the World Wide Web.

HTTP Hypertext transfer protocol, the underlying protocol used by the World Wide Web.

KB Kilobyte => 2^{10} bytes => 1,024 bytes.

Kbps Kilobits per second (file transfer rate). 1 Kbps is 1,000 bits per second.

MB Megabytes => 2^{20} bytes => $1,048,576$ bytes.

Mbps Megabits per second (file transfer rate). 1 Mbps is 1,000,000 bits per second.

Modem Short for modulator–demodulator, a device that enables a computer to transmit data over telephone lines.

Public domain software This type is software is not copyrighted; therefore users can modify the software but it cannot be sold for profit.

Secure transfer The communication is encrypted to ensure that intruders cannot interpret the data being transferred.

Server A computer running a specialized server component of client/server software to provide shared services. Some examples are mail server, FTP server, and Web server.

Shareware This type of software is available to try for approximately 45 days before purchase. If one wishes to continue using the software it must be purchased.

Spyware/adware This type of software secretly collects user information through the user's Internet connection for advertising purposes without the user's knowledge. Usually, this software is bundled as a hidden portion of freeware or shareware and installed automatically without the users' knowledge. Unknown to users spyware can gather user information, e.g., e-mail addresses, passwords, and credit card numbers.

Streaming A technology that allows listening to audio or watching videos without downloading the whole file.

URL Uniform resource locator, a globally unique address of documents and other resources on the World Wide Web.

CROSS REFERENCES

See *Client/Server Computing; File Types; HTML/XHTML (HyperText Markup Language/Extensible HyperText Markup Language); Internet Literacy; Internet Navigation (Basics, Services, and Portals); TCP/IP Suite.*

REFERENCES

Worley, B. (2002, January 4). Tech Live. Retrieved January 30, 2002, from http://www.techtv.com/news/security/story/0,24195,3366715,00.html

TuCows (n.d.). Retrieved November 1, 2001, from http://www.tucows.com

FURTHER READING

Bandwidth Place (n.d.). Retrieved November 1, 2001, from http://bandwidthplace.com/speedtest

CD-ROM FTP (n.d.). Retrieved October 20, 2001, from ftp://ftp.cdrom.com

CNET (n.d.). Retrieved October 25, 2002, from http://www.cnet.com/

CNET's Download (n.d.). Retrieved October 20, 2001, from http://download.com

CNET's Shareware (n.d.). Retrieved October 20, 2001, from http://shareware.cnet.com

Compression (n.d.). Retrieved May 19, 2002, from http://www.davesite.com/computers/compress.shtml

Gnome's Downloading (n.d.). Retrieved October 20, 2001, from http://download.gnome.org

Newbie Club (n.d.). Retrieved October 20, 2001, from http://newbieclub.com/download

PalmGear (n.d.). Retrieved November 20, 2002, from http://www.palmgear.com

Version Tracker (n.d.). Retrieved November 20, 2002, from http://www.VersionTracker.com

Webopedia's Online Encyclopedia (n.d.). Retrieved November 20, 2002, from http://www.webopedia.com/quick_ref/fileextensionsfull.asp

E

E-business ROI Simulations

Edwin E. Lewis, Jr., *The Johns Hopkins University*

INTRODUCTION

Since the 1980s, U.S. corporations have shown steady improvements in efficiency. This is the result of competitive pressures and the desire to optimize investor returns by executive management. As reported by the Bureau of Labor Statistics (U.S. Department of Labor, 2001), U.S. worker productivity continues to improve and the existing operational processes used by those workers continue to be optimized in most firms to support continued growth. Worldwide, corporations of other nationalities are driven by similar competitive influences originating from either local or regional firms or from international competitors, such as those in the United States. Successful firms approach competition in several different ways but in most cases, executive management responds by focusing its efforts on developing strategies that will increase revenues, reduce costs, and make improvements in existing processes. In striving to support these activities, many firms have come to the conclusion that the introduction of one or more e-business solutions into various operational areas of their organizations offers a primary means to implement these strategies.

Forrester Research predicted that corporations would spend approximately 3% of their revenues on e-business procurements in FY2002 and that this would continue to bring the financial costs of these types of systems into focus (Surmacz, 2001). This 3% investment may realistically equate to 40–60% of a firm's entire capital assets budget. Even at this funding level, not every project can be funded and there is no guarantee that those that are funded will generate the anticipated benefits. Thus, the projected benefits of capital expenditure on an e-business solution are legitimately brought under scrutiny by executive management. The most effective companies are beginning to measure a mix of business objectives, costs, and benefits to validate strategic implementation of e-business solutions. They then hold managers accountable for the results that they achieve during the implementation process and subsequent operation of the solution (Colkin, 2002).

For clarification, the term "capital investment" (or "expenditure") refers to the amount of money allocated, consumed, or used during a particular period (event horizon) to acquire or improve long-term assets such as property, plant facilities, equipment, or resources generically identified as capital assets. In this context, expenditures on research and development projects or projects of limited life would not be included in this definition. Depending on their accounting practices, some firms identify e-business solutions or larger IT systems, such as enterprise-wide systems, as capital assets because of their value to the corporation, their expected productive life, and the expenditure in capital to build, deploy, and maintain those systems. Accordingly, the purchase of these assets is referred to as capital expenditure. In all cases, the basis for any investment (expenditure of money) in capital assets must reside in sound business planning, performance, and measurable returns based on a quantifiable methodology. The challenge for any corporation's executive management becomes selecting which investment opportunity will offer the greatest benefits to the corporation over the current event horizon.

The best way to accomplish this selection between the different investment opportunities is to calculate each solution's or asset's estimated costs and benefits and then compare all available solutions to one another. This approach requires that consistent methodologies and processes for creating those estimates be developed and strictly followed. Values and risks that cannot be quantified may be considered by executive management as part of their decision-making process for determining capital expenditures, but from a quantitative valuation process, they should be excluded from the simulations and models. The cumulative results of the consistent application of valuation practices should be the identification of one or more investment opportunities that will provide the greatest returns to the corporation as compared to all other investment opportunities that are available. In general terms, this capital expenditure planning process as it relates to the procurement of capital assets is referred to herein as capital expenditure (CE) modeling or simulation.

Investments in information technology (IT), from a capital expenditure perspective, should be considered like any other capital-intensive investment by a corporation. IT investments usually have comparatively shorter life expectancies than other types of capital projects. A facility may have a useful operational life of approximately 25 years, whereas an e-business solution's hardware components may have a useful life of only 5 years. However, the intellectual property value of the developed software may be substantially longer. Often, e-business solutions require more technical labor to build and then subsequently to run the solution than other projects. Additionally, to maintain a valid operational environment, numerous upgrades to the deployed system will be required on an ongoing basis. From a commonality perspective on financial resource consumption, both IT and non-IT assets have similar process characteristics. Executive managers make the decision to invest in capital assets. The expenditure of funds is authorized and expensed and the resulting capital asset is depreciated over the life of the system or equipment.

IT investments, because of the percentage they make up of the overall capital investment budgets of most corporations, have become outcome-based with return on investment (ROI) usually being considered the key element in the decision-making process. The U.S. General Services Administration (GSA Budget 2002) Fiscal Year 2003 Budget Overview, which offered a projected $6.5B IT budget for that year, put it succinctly in their Capital Planning Guide (GSA Guide 2002), by stating that "IT projects have to clearly communicate their payoff to justify the investments. . . . " It is important to accurately predict the amount of capital expenditure required and the expected benefits that will be derived from that expenditure. To establish a standard point of analytical reference, both cost and benefit are measured in financial terms.

In addition to the cost estimates associated with an IT-based capital asset investment opportunity, estimates of its future benefit to the corporation are required. Benefits should be identified and used on the condition that they can be quantified with a level of reasonable confidence. They can range from easily identifiable cost savings, such as those brought about by staff reductions, to benefits that are much more difficult to quantify, such as projected revenue increases from a new point of sale business-to-consumer site. In these and other cases, benefits will be incurred but the amounts and duration are hard to determine. Ultimately, the simulation will need to be structured to include benefit estimations that will vary over time.

It is important to recognize that an organization's CE simulations used for e-business solution predictive or *ex ante* analysis are corporate assets of significant value. The value of a firm's simulation is directly proportional to its ability to accurately and successfully predict a specific solution's financial viability over a range of economic and environmental variables. Each firm strives to ensure that its simulation will provide better investment guidance than its competitors' simulations, giving them a true competitive advantage. As most firms recognize, once developed, tested, and validated, these simulations become, in effect, trade secrets and their dissemination is restricted as with most other intellectual property. This effectively limits the availability of complete models and simulations for download off of the Web. This fact also tends to limit the distribution of IT CE case studies based upon data obtained from operational firms.

E-business Systems

The generic term *e-business systems*, as used in this context, identifies IT systems that use Internet-based technologies to facilitate business transactions. At the highest level, the vast majority of e-business systems use Web services (essentially services provided or "served" by Web servers) to deliver content and exchange data with individual users or with other systems. This is not to say that other, non-Web enabled systems or services would not be included within an overall e-business solution. On the contrary, any solution of this type would almost always incorporate other communications systems and application services as well. However, an e-business solution cannot be classified as part of this category without a Web services component. Using this definition, most enterprise-wide systems, such as enterprise resource planning, enterprise relationship management, and customer relations management, provide some form of content delivery and systems management through a Web services interface.

This characterization is important because of the environmental and operational issues often associated with the design, development, operation, and support of systems delivering Web services. In almost every case, the time between recognizing a need and deploying an e-business solution is compressed. This results in many issues ranging from design integration to operational support problems. Often, even fundamental components of project management and IT system design are not fully examined before a project is pushed forward, investment budgets allocated, and costs expensed.

It is also important to recognize that e-business solutions are more complex than many other types of IT systems. They include component, system, and application integration activities, all of which are complex themselves and may directly affect one another. Regardless of what the various vendors would represent in their sales literature, much of the technology used within e-business solutions is often new and, accordingly, may not be mature beyond a very limited set of parameters. After examining a specific product, the analysts may agree with one another that its potential to benefit a corporation is significant. The subsequent implementation into that corporation's operational environment may identify a product's lack of maturity. Examples of this are common across the survivors of the Internet implosion. A common example was found in the initial releases of dynamic content Web server engines. Some products would work exceptionally well on Solaris platforms, whereas the same system running on an HP-UX platform would require significant tuning or patching. With other systems, the opposite was equally true. Unfortunately, the discovery process to identify and then to resolve these issues requires significant resources and time, resulting in increased costs and delays in realizing returns on the investment. E-business solutions are specifically designed to provide benefits by creating operating efficiencies, reducing costs, or increasing customer satisfaction. The three generic classes of e-business solutions are server-to-server, server-to-client, and client-to-client. It should be noted that these classes are not mutually exclusive and may be combined within the same e-business environment, resulting in a very capable system.

The challenges of the integration of e-business solutions into other systems can be significant. In most cases, effective e-business solutions include connections to one or more external systems to support data transactions. The configuration of remote systems may be unknown or uncontrollable, changing from time to time, yet functionality between the environments must be maintained and supported. This may include connections to remote systems used by other organizations or simply consist of developing code that will work with different versions of Web browsers. The number of users, operational loading, and performance parameters may change rapidly, possibly based on a market event or the effectiveness of a simple advertising campaign.

The content of an e-business solution is seldom static and requires either automatic or manual updates, often requiring different types of staff to create and then install these updates. Many of the most effective e-business solutions utilize a firm's back-office systems as a source of deliverable content. This requires connectivity to these various back-office or legacy systems, often requiring different middleware applications to support the different connections. By extension, it also requires enhanced and layered security that may not be common practice for many organizations.

Finally, adding to the complexity, most e-business solutions use a combination of technologies. Though these technologies may not be unique, the combination of them adds complexity to the solution's implementation and operation. Most IT staffs specialize in a set of standard technologies. If the designers, developers, implementers, and operators are not familiar with the technologies that are to be used in combination within this solution, design requirements will be missed, build errors will occur, and incorrect configurations will be made, adding cost and risk to the IT capital asset deployment that is an e-business solution.

E-business Solution Cost and Benefit Estimation

Mitigating the risk from these and other problems will add costs to a project. The primary means of mitigation is to allocate more time and/or skilled resources to advanced preparation and planning in the initial stages of the project. This requires the expenditure of additional labor and its associated costs in an accelerated delivery and budget conscious environment. Many firms resist these initial phase investments and accordingly, many projects fail to reach their full potential. Those organizations that fail to recognize the potential complexity of an e-business solution deployment and the need for additional planning to support it almost always incur unanticipated costs while they learn to integrate and subsequently use these systems in daily operations.

Even with the capable tools to supplement their estimating efforts, many firms simply do not accurately perform cost and benefit analyses nor do they interpret the resulting data correctly. Jupiter Research reports that most firms that perform ROI analysis on their IT projects do so with staffs that provide either an inherent bias to their analysis or that have not been properly trained to fully recognize the many aspects of this type of analysis (Mello, 2001). Staffed assigned from the Information Technology departments to perform an analysis may understand the technical details of their projects but seldom have access to information regarding the financial impacts of their projects on the company. Conversely, finance and accounting departments can determine the impacts of the expended costs and perform detailed *ex post* (or historical) evaluations of a project if the needed data is collected but seldom have complete or sufficient data for accurate *ex ante* evaluations. The key mitigation of this potential failing is training individuals in the use of the tools and methodologies that are provided to them for estimating purposes. This includes estimate preparation as well as simulation output interpretation.

The process that facilitates the collection of estimates must recognize that both the benefits and the costs contain many tangible and intangible elements. The detailed quantification of those elements is often difficult to accomplish. Intangible costs or benefits should be considered carefully before they are included as quantified values within a simulation. E-business solution analysis requires the creation of a simulation that reflects an organization's approach to business, the quantification of many variables, both known and estimated, and a use methodology that provides consistent valuations between solutions.

Simulation Design Components and Methodologies

> "One accurate measurement is worth more than
> a thousand expert opinions."
> —Admiral Grace Hopper, U.S. Navy

CE simulations utilize formulas and variable data that mimic objects, processes, and activities found in the real world. In developing a simulation, several goals for the basic design should be established first. These goals are neutrality in the simulation's formulation and structure, repeatability in its output, and reliability in its results.

1. Neutrality in the formulation and structure of a simulation refers to avoiding a bias that unduly penalizes

one variable element over another. If allowed to be introduced and exist within the simulation, such bias will always skew its results.

2. Repeatability in the output of a simulation requires a consistent and stable basis for the identification and use of variables. Variable elements will change over time, but a valid comparative analysis requires consistent use where possible.
3. Reliability in the results generated by a simulation remains one of the most critical elements of any predictive analysis. A simulation can perform flawlessly from a mathematical perspective but the subsequent results must closely reflect the actual results generated by the solution being analyzed.

The purpose of the simulation is to generate reliable and consistent results on which optimal investment decisions can be made.

Simulation Development

Once the overall goals are recognized and established, the specific approach to the simulation's design may be addressed. The basic design criteria require that the simulation be able to accurately predict each solution's financial costs and benefits while factoring into those predictions the organization's future environmental state. The recommended design concept is that the core elements of the simulation are standardized. These elements would include all base formulas, such as NPV, IRR, or ROI, and environmental factors, such as corporate interest rates or facility costs. This approach will result in any project using the simulation being subject to the same valuation formulas and environmental constraints or expectations. Solution-specific variables would remain non-standard for each analysis. Nonstandard variables would include items such as software development and hardware platform costs. With this approach, the unique aspects of each project can be measured, allowing a more efficient base for comparison purposes.

The environmental state of a firm refers to external financial and economic variables that are considered to be outside of the control of the organization. These variables impact all organizations within that same market segment or within the same organization or division within that organization. They may include items such as interest rates, inflation, and other economic factors in general. Industry-specific issues such as raw material costs or skilled labor costs within a geographic region also fall into this category. The valuation of these factors are usually the result of the corporation's financial structure and its accounting practices. These are usually incorporated into the simulations as items such as general and administrative costs and fringe benefits as applied to labor costs.

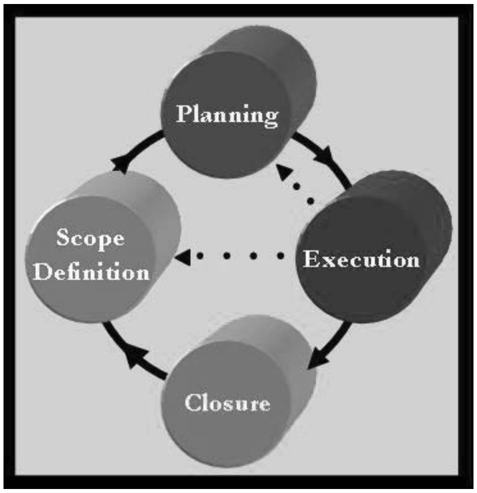


Figure 1: Project Lifecycle. (Art by Holly Robedeau.)

The simulation itself is one of several tools and processes that are used by most corporations for performing CE analysis. Estimating tools should include design engineering models that predict lines of error-free code based on developer experience and the use of processes that ensure consistent tool use in estimating each project. Though the simulation utilizes solution-specific estimates as variables, the acquisition and the quality of those variables, from a timeliness and accuracy basis, must also be addressed. Estimating processes vary from firm to firm but the most effective ones rely on a consistent methodology that is frequently validated for accuracy.

Optimally, during the final phases of a project during closeout, a "lessons learned" is also performed where errors, omissions, inaccurate estimates, and oversights are identified. The identification process is usually supplemented by an *ex post* project valuation activity that will identify these elements if performed properly. Where appropriate, the associated processes and simulation elements that were used to create the original estimates are modified to mitigate a repeat of the previous error. Training is usually associated with this type of activity to make sure the estimating teams are kept appraised of historical problems. Changes resulting from "lessons learned" analysis are designed to correct problems to an existing model and not to make core or fundamental changes to it unless absolutely necessary. They are then incorporated into the simulation or estimating process for use in subsequent modeling exercises.

Approaches to E-business Simulations

Most firms' methodologies take one of two general approaches towards developing the estimates for use in CE simulations. These are either the framework or the project-centric data acquisition methodology. A framework-based approach is structured to consider enterprise-wide projects (those that are very large in scope and affect multiple departments) and focus on process optimization to assure efficiency and reduce costs. These projects are designed to identify, reorient, and baseline departmental processes across the organization through a series of sequential steps. Initially, they seek to eliminate redundant roles and responsibilities between departments and baseline each department's responsibilities. Once completed, they will then seek to identify the technological solutions needed to optimize those recently baselined processes. A project-based approach tends to view the implementation of IT projects from a technical perspective, with a limited focus on process re-engineering beyond the processes that relate to the immediate system.

Framework-Based Financial Simulations

A framework-based financial simulation is an extension of a process developed to assess a firm's overall information technology framework architecture. This approach was created to mitigate the problems encountered as firms began to undertake the implementation of very large and complex IT projects, such as enterprise resource planning implementations. These firms discovered that the scope of those types of activities was too great to be managed within a reasonable amount of time using a normal project-based approach in most circumstances. To be successful in the implementations of these types of projects, any underlying architectural problems associated with roles, responsibilities, and processes within the various departments or organizations within a firm needed to be identified and corrected. This was followed by a comprehensive analysis of each department's business processes and their associated IT architecture. The information derived from this analysis enabled the corporation to begin the selection process of an application, or series of applications, to add efficiencies to each organization within the corporation. The Zachman Framework, developed by John Zachman in 1987, is usually considered the premier approach to this type of architectural modeling (Hay, 2000). Once an application or series of applications is selected, a further review of business processes is required to adapt to or accomodate the use of those applications; revised staffing profiles will need to be estimated and training and application support costs examined. Framework estimating uses this process as a series of data collection points and seeks to determine implementation costs and benefits using a comprehensive *ex ante* analysis methodology.

Each application and department is subjected to a cost–benefit analysis considering as many costs and benefits as are identifiable within the time available. The main problem with this approach is the collection of a very large amount of data from many sources and then preparing accurate estimates. This process is then followed by executing the project tasks against those estimates before the data change. In all cases, the data on which the estimates are based will change over time. Organizations within a firm do not stop operations while potential impacts are assessed. Baseline core competencies are subject to change for many legitimate operational reasons. The more departments, divisions, and organizations there are within a firm, the greater the geographical dispersion of those organizations, and the longer the initial assessments take, the greater the risk that overall estimated costs and benefits will become erroneous before they are compiled and an expenditure decision is made. This reality requires significant effort to be expended quickly. The implementation of an Enterprise Wide System may take more than 36 months and cost in the tens of millions of dollars. Accurate estimates to address capital allocation, cost of money and related factors are critical to the viability of the project.

Project-Based Simulations

A project-based simulation is developed to predict the operational value of a project that was created using a set project development methodology. As defined by the Gartner Group, a project methodology has four primary phases. These are the scope and requirements definition phase, the planning phase, the execution phase (which includes monitoring and managing the performance of the deliverables of the project), and project closeout. The Project Management Institute classifies project phases similarly. (see Figure 1.)

The initial data collection activities that are performed to quantify variables for use within CE simulations is usually performed during the first two phases in a project, the scope and requirements definition phase and the planning phase. In many corporate project models, detailed estimates are first prepared in the beginning of the execution phase (phase 3), which implies that the decisions relating to capital asset expenditures have already been made by management. In working with e-business solutions, the detailed estimates should be included within the first two phases of a project if at all possible, well before a decision is made and funds are obligated. As many as 50% of all IT projects overrun their initial budgets. Errors greater than 80% over the initial budget are common (Shulte, 2001). This is unacceptable by any standard and usually attributable to poor estimation in the early stages of the project when the initial funding was requested or obligated. The mitigation for this is to prepare detailed estimates during the planning stage.

The project-based simulation should have the ability to incorporate an extensive series of solution-specific variables as estimates. For accurate predictions of a solution's value to be generated by a simulation, the maximum number of variables should be identified and quantified to the fullest extent possible. Additionally, these variables should be identified to a level that documents their numerous interrelationships and interdependencies. This may be accomplished from within the simulation using mathematical permutations or by using simple raw data manipulations as a data input strategy. In support of this, the estimation methodology should deliver variable data estimates that can be used by the simulation without manipulation and that readily fit the simulation's data entry structure. This avoids translation errors when the formats of the simulation and the estimates do not match. A simple example of this is for the firm to establish a methodology where labor estimates are recorded using tenths of an hour as the measurement scale. Estimates using minutes as the measurement scale would be rejected and reworked to reflect the appropriate standard.

CE SIMULATION DESIGN ARCHITECTURE

In most corporations, capital asset projects requiring capital expenditures in amounts greater than certain monetary thresholds are subject to some form of screening. To ensure that each project is fairly reviewed in comparison to other projects, it must be remembered that all projects should be planned and estimated using a consistent set of tools and methodologies. Once the estimated costs and benefits are identified, a similarly consistent methodology should also be followed to compare opportunity to opportunity. The goal is to determine the valuation of each project so that accurate comparisons between projects can be made. The design architecture of the basic simulation is critical to this approach.

A successful design for a CE Simulation is built around a set of formulas and environmental variables that have been recognized as financial standards by the corporation. These elements are combined into a core simulation that is validated for mathematical functionality and accuracy. The design will allow the core simulation to utilize operational variables to determine specific project valuation. Once the design and resulting simulation are built, the simulation is tested and validated. Simulations that are suitable for e-business project analysis use the same base components as any other CE model but require the ability to incorporate a significantly larger number of operational variables than are required for other, non-IT projects.

Simulation Standardization

In developing a simulation, there are several aspects of the standardized formulas that must be addressed. The initial step in this process consists of identifying the values of current and future environmental variables that will affect the solution's valuation and that will be used by those formulas. By extension, this requires that the firm agree internally, across organizations and departments, on a standard approach to determining these values. The actual valuation is usually done by a group within the firm's finance or accounting department and released through them to the rest of the firm, thereby meeting these requirements. These environmental variable valuations are based upon a series of known or probable events (Arsham, 2002).

Some of the key environmental variables that many firms consider suitable for standardization include

1. **Interest Rates.** The cost of borrowing money, or the interest that could be earned by depositing money over time, is a primary financial element suitable for standardization. In that many interest rates and sources for capital that are credit- and institution-based, corporate standardization of these values is needed. As interest rates change over time, the cost of money should be identified for use over time as well.
2. **Inflation.** The annualized rise in prices that is often found in market-driven economies is of concern in any financial simulation. Inflation occurs when the demand for goods and services exceeds the available supply, increasing the price for those goods and services. Inflation may be associated with individual labor categories or commodities.
3. **Labor Rates.** There are several ways to present labor rates in any financial model. These could include using the actual rate paid to a specific employee who will do the work, using category rates (the average or weighted average rate for everyone with the same general skill set who performs similar work at a specific location for a company), or using departmental rates (the average or weighted average rate for everyone within a department, including all administrative and managerial employees).
4. **General and Administrative Costs.** Labor costs usually include some benefit costs, management costs, and HR costs, as well as a tax burden associated with each rate. These costs usually are considered to be indirect costs and may vary with location within the corporation. In larger firms, these costs are also subject to frequent changes as reorganizations and normal staffing changes occur. For these reasons, it is usually recommended that G&A costs be excluded from all *internal* analysis. However, for *external* pricing purposes, such as that used in a proposal, G&A costs are usually considered to be legitimate, reimbursable costs and should be included within these types of projections.
5. **Overhead, Facilities Costs, Material Handling, and Administrative Charges.** Depending on the firm, some or all of these elements may be charged to individual projects. In estimating costs, recognize that these are cost factors that could significantly impact a tight project budget. Additionally, these costs can change over time as a result in changes in accounting practices. It is not uncommon for these costs to be excluded from use by some simulations.

Other environmental state variables may be considered based on industry-specific standards. Additionally, the values assigned to these variables will change over time and it is imperative that this change be reflected in any simulation. This list is not meant to be inclusive of all possibilities; it represents a standard set of elements that may be recognized within most models and simulations.

Once the standard variables are identified, a set of formulas and their interrelationships within the simulation can be created. The selection of formulas will be at least partially industry-specific. For example, a firm that utilizes direct buys of imported steel for manufacturing may seek to develop simulations that include currency risk management hedges and internal rate of return calculations for estimation of its manufacturing process costs. These simulation components would provide little or no value for a consulting firm that provides accounting audits only in the United States.

For representative purposes, the following formulas should be considered for inclusion into most basic comprehensive simulations:

Net Present Value (NPV)

This calculation will present a valuation of the investment in future years:

$$\left[\left(\frac{CF1}{(1+r)^1}\right)+\left(\frac{CF2}{(1+r)^2}\right)+\left(\frac{CF3}{(1+r)^3}\right) + \cdots\cdots + \left(\frac{CFn}{(1+r)^n}\right)\right] - I = NPV$$

Solve for NPV. *CF* equals estimated cash flow, *r* equals the current investment rate, *n* equals the period over which the net present value is calculated, and *I* equates to the initial investment. n is expressed in a manner consistent with r (e.g., if r is annualized, then n is in years). If NPV > 0, then an investment can be made with a positive return expected. Conversely, if NPV < 0, then the investment should be rejected in most cases (Office of Secretary of Defense, 2000).

Internal Rate of Return

The internal rate of return (IRR) is used to measure the interest rate at which the present value of a series of investments is equal to the present value of the returns on those investments. The IRR for a particular solution is usually compared to the overall IRR for an organization or corporation.

$$\left[C0 + \left(\frac{CF1 - CST1}{(1+IRR)^1}\right) + \left(\frac{CF2 - CST2}{(1+IRR)^2}\right) + \left(\frac{CF3 - CST3}{(1+IRR)^3}\right) + \cdots\cdots + \left(\frac{CFn - CSTn}{(1+IRR)^n}\right)\right] = 0$$

Solve for *IRR*. *CF* equals estimated cash flow for that period period, and *CST* equals the estimated cost for that period. It should be noted that both cash and cost would change annually. *n* equals the period over which the *IRR* is calculated and is expressed in a manner consistent with *IRR*. (If *IRR* is annualized, then n is in years.) The results of this formula are compared to desired or expected returns for any project based on stated corporate standards (Wu, 2000). If the returns are greater, the project should be accepted. If they are less, then it should be rejected.

Payback Period

$$Payback\ Period = \frac{INV}{(CF1 + CF2 + CF3 \cdots CFn)}$$

Solve for Payback, where *INV* refers to the initial investment and CF equals estimated cash flow over time. The answer is presented as a factor of the time applied to the cash flow measurement (Wu, 2000).

Return on Investment

$$ROI = \frac{(CF1 + CF2 + CF3 \cdots CFn)}{INV}$$

Solve for ROI, where CF equals estimated cash flow over time and *INV* refers to the initial investment (Wu, 2000). ROI is presented as a percentage value either above or below the initial investment. If the ROI percentage is greater than 100%, then it is probably a financially justifiable project. If the calculated ROI is less than 100%, then the investment dollars could be better spent elsewhere.

$$0 = CF_0 \frac{CF_1}{(1+r)^1} + \frac{CF_2}{(1+r)^2} + \frac{CF_3}{(1+r)^3} + \frac{CF_n}{(1+r)^n}$$

Other formulas should not be excluded from this list and are, as with the environmental state variables, dependent on industry-specific requirements or standards. Additionally, many firms use variations of these formulas as part of their accounting and finance systems. Those formulas should be given preference over the examples offered here. The source and weighting of the values used in these formulas should also be modified as needed. It should be noted that in the formulas containing *INV* (for investment) as an operational variable, this value specifically refers to the expected funding expenditures. In most scenarios, changes to this variable will have the greatest impact to the overall results of the analysis and, by extension, the viability of the project. Additionally, this is also the one variable that can readily be managed by the corporation and the project management staff.

OPERATIONAL COST AND BENEFIT ESTIMATES

The number of operational cost estimates within an e-business solution can be significant and complex, involving many more variables than found in other capital asset projects. The installation of a Web server, as part of an overall e-business solution, must be optimized prior to use. This effort requires several specialized skills and the effort needed to be expended can be considerable. A systems administrator will need to set up the hardware, install the operating system, and then install and configure server applications. A security specialist may be needed to apply security patches to the OS and application. Additional build issues could include applying application-specific load balancing configurations, multihoming the servers, installing and configuring database and application agents, creating and installing custom code, load-testing the code, and performing session state and overall performance optimization tuning. In most cases, this is only one of many servers in any solution including application, database and special purpose devices. Similar work will need to be done with the networks, back-end system integration and content development applications, and management components.

Each of these solution components has many discreet cost estimates associated with it and often, each is unique to the solution. The identification of the discreet labor costs and all related details associated with these elements is critical to achieving accurate estimates that can be used as a basis for project comparisons. Complete and detailed estimates will lower the probability of overruns if the project is authorized and fully funded.

Benefits often prove more difficult to accurately estimate than project costs. The identification of some benefits, such as those derived from the elimination of positions that are automated by a particular solution, is straightforward. Others may require the use of their own predictive models and estimates, the creation of department specific surveys and establishing satisfaction criteria for value quantification. For a business-to-consumer (B2C) site, out-year revenue estimates may reflect significant uncertainty. A business-to-business (B2B) solution, such as those created for supply chain management purposes, may be dependent on the commitment of suppliers to continue to support the solution. For benefit estimation, a conservative approach and one that recognizes tangible benefit values first is recommended.

From strictly an analytical perspective, the quantification of savings is usually more accurate than projected revenues or the future value of customer satisfaction and good will. Though, at times, these can be reduced to a value or range of values and they are extremely important, an accurate mathematical valuation is difficult to obtain.

Benefits are as divergent as the e-business solutions from which they are derived and are always project-specific in their determination and quantification.

Regardless of the mathematical validity of a simulation and the consistent use of standard variables from project to project, the data generated by the simulation remain only as good as the operational variables that were estimated. Cost and benefit data are project-centric in their collection and presentation. They must be identified through a specific and detailed approach that is consistently used. It is generally assumed that using the same simulation in a consistent manner would be a standard practice in most organizations but, unfortunately, using different simulations for different project types is commonplace (Remenyi, 2000). This must be avoided wherever possible. The analytical approaches should always be as similar and comprehensive as possible. Projects involving IT solutions may have significantly more variables than others and, accordingly, the models used should be able to accommodate those additional variables.

Operational Cost Identification: The Framework Methodology

Framework-based estimating methodologies, used in support of division- or corporate-wide systems implementations in medium-sized to large organizations, are probably the most difficult to develop and use. A framework approach usually follows a series of general steps, each of which may be customized to support each corporate environment.

The initial step in this process involves identifying the roles and responsibilities of each department within the organization. These are examined from the perspective of each department's original charter and compared to its current activities. On review of each department, non-core functions that they are currently performing are re-assigned to the departments responsible for those functions as part of their core responsibilities. From a cost efficiency perspective, the duplication of services across departments is one of the most common problems in any organization. Departments commonly create their own internal non-core services or support infrastructure in response to their inability to receive those same services or information from the responsible department within their firm. As an example, there is usually a person in most departments whom co-workers go to for PC help instead of first calling the help desk; in other cases, management assigns someone to keep a spreadsheet to track the hours that are spent by a department on a project instead of getting that information for the accounting department. These are often referred to as "ghost" services. By creating their own support, departments meet their immediate needs but become increasingly cost and operationally inefficient over time. Creating and delivering internally provided services utilize departmental resources that should be spent on the department's core mission. With the reallocation of core roles and responsibilities as a result of a Framework analysis, revised departmental responsibilities and updated budgets are usually established to reflect the current departmental configurations of the corporation. These revised budgets are compiled and the resulting figure used as a baseline against which future project costs and benefits are compared.

At the conclusion of this initial activity, roles and responsibilities have been optimized within the existing environment. The next major step is to review the business processes of each department. Once the processes are known, comparisons between departments can be made. This activity will seek out redundant activities between departments, such as entering the same data into two different systems or performing manual transactions where automated transactions would be more efficient. As part of this comparison process, various IT solutions would be examined that will solve as many of the discovered problems as possible.

At the conclusion of the initial process assessment phase, the process of selecting an enterprise-wide system or, at a minimum, one or more e-business solutions can begin. This process includes system selection and costing. Enterprise-wide system application selection almost always requires outside expertise. Regardless of the products or modules selected, new labor, hardware, software, integration, and support costs will all be incurred by the corporation. Accordingly, the impact of the applications on each department must be identified and their costs and benefits estimated. As with all other aspects of a framework approach, these impacts must be estimated for each department and then combined into one integrated estimate package. The scope of this effort is significant in almost every respect. A detailed process and activity level analysis cannot be done without incurring significant costs over long periods of time. The combination of the departmental reviews, process reviews and application estimating often takes many months.

Some firms, looking for alternatives to lengthy data-gathering activities, attempt to use a business case analysis approach to justification for adapting large system deployments (Stien, 1999). This approach generally sets forth a series of well-founded arguments for the implementation of an integrated solution but fails to provide a sound financial basis for justifying the investment. An approach that seeks to identify the required expenditures and expected returns is more appropriate.

The reality is that, as the many elements within the overall framework assessment are identified and the magnitude of having to estimate the costs and benefits becomes known, a detailed quantification approach using internal resources is usually replaced by a more expedient approach. The solutions available are either the creation of the required estimates by the department managers or the use of an external consulting firm to perform some or all of the estimating functions.

Internally developed estimates will be prepared by department managers and their teams. The problem with this is that the level of effort is usually significant and reduces the ability of the department to perform its normal work assignments. Internal staff members seldom have the experience needed to perform process re-engineering estimating, nor do they have the key metric data needed to validate their estimates. Additionally, their assessments will reflect some inherent bias as they review their own processes. It must be recognized that if a department assumed a role or responsibility of another department, it did so to satisfy a need that was not being met elsewhere.

Facing these issues, corporations often consider experienced outside consulting firms as the preferred solution, believing that they will provide a greater level of accuracy and will be less disruptive of normal operations during the estimating phase of the activity. These firms can provide qualitative estimates using historical data obtained from previous work in firms of similar size and structure and in the same market segment. The firms that offer these types of services do so from highly specialized practices. There are only a handful of firms that have the experience and historical data to provide the level of service needed to support a large enterprise-wide system deployment. IBM Global Consulting, CSC, and EDS are leaders in this area and all have consulting practices that provide these services.

During and after their previous implementations, these consulting firms would have compiled extensive metrics on the incurred costs and actual benefits that were realized by their clients. These sets of metrics, validated by *ex post* analysis, are used as a baseline for comparative analysis on new projects. As a point of reference, the right of the consulting firm to collect metric data and use them on subsequent engagements for comparative purposes, while maintaining source confidentiality, is usually contained within the consulting firm's engagement contract. The sophisticated use of metric data for framework estimate analysis is one of the primary products offered by these types of consulting firms. They are able to provide accurate estimates because of the sophistication of their simulations and their use of these metrics. Within these parameters, the consulting firm is usually tasked with performing the framework architectural analysis and the subsequent financial analysis. This would include process re-engineering services and recommending the optimum technical solution. The general methodology that is usually followed is for the consulting firm to normalize the current implementation estimates to similar metrics that they have on file. Once this is completed, a relatively accurate set of cost and benefit assessments can be created. The level of detail will usually be less specific than that found in a project-centric simulation. The end result of this activity is that the values documented as the result of the framework analysis process can be compared to other capital asset investment opportunities. Executive management will then be in position to compare the consulting firm's solution valuation with the valuations of other capital investment opportunities.

It remains critical for the firm to select the appropriate consulting contractor and the optimal technical solution and of equal importance, to manage this effort correctly. There is a significant body of knowledge available on this process. Though using a contractor for assessment is expensive, it usually remains a cost-effective alternative to performing a predictive assessment on a department-by-department basis using internal resources. For a firm attempting to assess potential costs and benefits prior to the initiation of enterprise-wide implementation, this approach is usually the most effective.

Operational Cost Identification: The Project-Centric Methodology

Project-based estimating methodologies are commonly used for the analysis of e-business projects. They are effective in analyzing most projects and are familiar to most managers. This type of simulation seeks to accurately identify as many costs and benefits as possible for each solution. To accomplish this, several conventional project management techniques should be employed. First, the overall scope of the project is broken down into a series of work-flow processes. Each work-flow process consists of a series of sequential and interrelated work-flow elements such as

1. Designing the solution's software,
2. Building the software,
3. Testing the software, etc.

Within each of these elements, there are a series of hierarchical tasks, subtasks, and activities, with activities being the smallest work component. These are then compiled into a work list. Within each of these elements, the identification of what needs to be done to complete the overall work element, the work schedule, and the resources that are required to perform the work will need to be estimated.

A recommended methodology for managing the results of breaking down the work-flow process is the use of a work breakdown structure (WBS). A WBS is a sequential numbering of the work list that also reflects the hierarchical structure within each work flow. As a guide, each element within the WBS is single-purposed and independent (but may have dependencies for starting or stopping), and it must be of a specific duration, have clearly understood deliverables, and, if needed, be created from detailed subtask elements (Berg, 2000).

In that the resource requirements for each work-flow element within the WBS must be estimated, it is also recommended that a formal process be created to consistently document and track each estimate. This is usually done through the use of a basis of estimate (BOE) form that identifies each component of the estimate in detail. Ideally, the technical estimator will document exactly, and in great detail, what he or she will need to do to complete the work, the labor by category, and all material resources that are needed, as well as an estimated duration to complete the work within each individual WBS element. Additionally, the BOE will also include dependencies and assumptions as part of the estimating process. With this documentation, estimated costs can readily be applied to each element within the work list, simplifying the capture of the overall operational cost components. Costing should be prepared by individuals other than those who prepared the technical BOEs. This serves as a check and balance within the costing process.

Of the problems with estimating operational cost variables, beyond the ability to identify all of the work that must be done and associate a cost with those elements, is determining the ability of the current IT staff to efficiently perform the required work. If an IT department is tasked with developing an estimated 350,000 lines of custom code for a supply chain management system, several aspects of that estimate should be taken into consideration within the estimating methodology. Identifying the labor resources that will be needed for the duration of the WBS element is critical. It is intuitive that the hours needed to perform this work will not be consumed at a constant rate (level loaded) for the entire duration of the project. The loading will change over time. If there are significant costs for this work, through the use of subcontractors or expensive programmers, the understanding of how these funds will be spent is also important for efficient money management. Accordingly, the simulation should have the ability to accommodate changes in resource productivity and allocation over the duration of the WBS element.

A common solution to this is the application of a productivity curve to the estimates provided on the BOE forms. Productivity curves approximate the distribution of labor over the projected work duration of the WBS element. Depending on the methodology developed, a standard set of curves is provided to the estimators that reflect different labor loadings (bell curve, ski-jump, etc.). The estimator then selects the curve that will most realistically reflect the labor staffing profile of that specific task. This provides the simulation with a standard mathematical structure to apply time value of money principles against individual tasks. Once the project is completed, an *ex post* analysis of the task will be performed and the curve selection validated or modified for use on future projects.

Within this context is the ability of the simulation to accommodate uncertainty within the estimates that are created. Accurate estimation of software development projects is notoriously difficult. There are many methodologies that may improve the accuracy of these estimates, such as the Constructive Cost Model, or COCOMO (University Southern California, 2002), will be used to increase the accuracy of the estimates reported in the BOE. For various reasons, they are not used by many firms and their own models, if they have them, are not as accurate. The result is that software development estimates often are inaccurate by a full order of magnitude (Porter, 2002). These errors continue to plague most e-business solution development activities. If there is legitimate uncertainty regarding the work to be performed, and other reasonable alternatives to risk mitigation are not available, then a weighting process should be followed. This would only be applied after the standard estimating process was completed and would clearly identify any weighting that was applied. The application of weighting should also be subjected to several constraints and used sparingly.

The first of these requires that the basic estimates identify the minimum resources needed to perform the work. The conscious padding of estimates by including additional labor or materials to mitigate risks must be eliminated from the resource estimation process. Only estimates based on reasonable and, to the greatest extent possible, justifiable expectations of costs should be weighted. The preparation of estimates in an uncertain environment is difficult at best and may need to be performed by the team's most senior and experienced members. Blue-sky or gut-feeling estimates should be avoided if possible and should be reduced to documented values to the greatest extent possible. Though uncertainty will still exist, its parameters will be available for the *ex post* analysis. This documentation process will establish a basis for improved estimates in the future.

Second, the individuals involved in preparing the resource estimates should be involved in the subsequent weighting process. Their understanding of the risk in this aspect of the project is the greatest. Finally, management must create an environment that will allow open discussions about the estimates themselves. A perceived or real negative reflection on the estimators will nullify the productivity of these discussions and must be avoided. Once these parameters are met, weighting can be identified within the cost estimation process and added formally through a visible process within the simulation. Without visibility, subsequent metrics acquisition activities that are associated with an *ex post* analysis of the WBS element will draw incorrect conclusions.

These estimating approaches are repeated for each work-flow process within the overall scope of the project. By so doing, the operational costs within each of these discrete packages can be effectively estimated. The risk of this approach is that the intangible aspects of a deployment, such as the repercussions of process re-engineering, may not be fully addressed. For a project-centric simulation, the recommended process is to quantify the details, validate the data, and then summarize each discrete estimate to calculate the total project cost. The results of this approach are definable costs leading to the development of a known capital expenditure requirement. Again, it must be emphasized that detailed estimates should be prepared before a comparison is performed and budgets are allocated.

Project-centric data gathering and estimation should be done using internal resources whenever possible. Ideally, they should be done by the people who will actually do the work. In most firms, a methodology can readily be developed and then optimized over time. Training for the estimating staff can be developed to support this approach and the returns on that training investment are immediate when the learned techniques are applied on the next project.

Simulation Testing and Validation

The testing of a capital expenditure simulation should be done at the individual component level, at the integration level, and as a complete system. It should be subjected to the same level of comprehensive testing as any other software application. The simulation should be tested mathematically and separately validated by comparing its results with known outcomes from other projects. With each iterative enhancement of the simulation, additional testing will need to be undertaken. Each series of tests should be comprehensive. Minor errors in programming can result in the commitment of significant capital resources to a less-than-optimal project.

CONCLUSION

The use of e-business simulations for analytical purposes can provide executive management with a sound financial basis for making investment decisions and optimizing capital expenditures. To accomplish this, the simulation may utilize the standardization of certain economic variables recognized across the corporation, the consistent use of the same set of simulations on all potential projects, and a dedicated *ex ante* variable quantification process (estimation). Regardless of the use of either a framework or a project-centric estimation methodology, a formal process must be followed and the individuals responsible for these activities trained on the nuances of each. As projects are evaluated and accepted, an *ex post* evaluation of the project's estimates and the simulation's manipulation of those estimates should be performed. During this process, oversights and erroneous estimates are identified and the root causes for any problems are sought out. The simulation model and estimating approach are then modified to correct these issues, providing a more accurate tool and process the next time it is used.

The steps in establishing the parameters for building a simulation that is suitable for any capital expenditure analysis should consist of the following best practices:

1. Establish a policy of consistent use of all simulations used for capital investment analysis. Unique simulations should not be created for individual project types, such as one for e-business solution analysis.
2. Establish a methodology for identifying environmental state variables. These should reflect the overall corporate expectations of the current and future values of each variable. Do not allow environmental state variables to be calculated on a project-to-project basis.
3. Establish a methodology for the utilization of industry-specific formulas in simulations that are recognized by the firm as a basis for performing financial analysis. If the formula is not relevant to the industry, do not introduce it into the simulation.
4. Use an established methodology for the identification of predicted costs and benefits. This may be either a project-centric or a framework-based estimation methodology.
5. Include functionality within the simulation to accommodate all cost and benefit estimates to the greatest extent possible within the methodology utilized. Most e-business solutions include many variables whose valuation should be included in the cost and benefit assessments. The exclusion of variables may degrade the accuracy of the simulation.
6. Recognize that project-specific cost and benefit data collection is a complex and time-consuming task. Due diligence requires a detailed, methodical approach to estimating. Allocate time in the estimation process to allow this to be done correctly. This should be done when initial solutions are being compared, not after authorization to proceed has been received.
7. Use minimum, reasonable estimates for both costs and benefits. Do not allow individual estimators to pad their estimates independently. A structured basis for estimation should be followed.
8. Include within the simulation a capability to account for uncertainty in the quantification of costs, benefits, and outcomes. This may be done either by individually weighting estimates or by applying a weight against the outputs of the simulation.
9. Aggressively test all aspects of the base simulation. Validate the results with actual data wherever possible. For each update or modification to the simulation, the entire testing process should be repeated.

The key to all of these aspects is for the analyst and the project team to pay very close attention to the many complex and interrelated details of an e-business solution. The technology is complex and expensive in terms of the procured items and the labor needed to make them work. The technologies may be new to many IT departments, so recent past experience may not be available. Understanding of that technology, diligence, and attention to detail are the keys to making capital expenditure simulations analyzing e-business solutions work as a decision-making tool for management.

GLOSSARY

***Ex ante* evaluation** A financial analysis based on predictive information relating to a project. This type of analysis is based upon concise and detailed estimates of the costs and benefits that will be generated during the early phases of a solution's life cycle. The primary tools for this analytical process are a structured simulation and a set of processes and methodologies that are designed to generate consistent, repeatable, and reliable results that can readily be used for comparison purposes.

***Ex post* evaluation** A financial analysis using actual project data compiled during the implementation of a project and subsequent operation of the solution. The primary purpose of this type of evaluation is to validate estimates and the accuracy of predictive models.

Web servers A generic term that refers to both a configured software application that serves Web applications or services and a physical hardware device. The leading software application Web servers are Apache, Internet Information Server, and Zeus, holding 56.38%, 31.96%, and 2.26%, respectively of the reported market (Netcraft, May 2002). Hardware Web servers refer to the computers or computer appliances that host the Web server software application. The term "Web server" is usually applied contextually.

CROSS REFERENCES

See *Electronic Commerce and Electronic Business; Prototyping; Return on Investment Analysis for E-business Projects; Risk Management in Internet-Based Software Projects.*

REFERENCES

Arsham, H. (2002, February 11). *Systems simulation: The shortest path from learning to applications*. Retrieved May 14, 2002, from http://ubmail.ubalt.edu/~harsham/simulation/sim.htm

Berg, C., & Colenso, K. (2000, April). *Work breakdown structure practice standard project—WBS vs. Activities PM Network* Retrieved June 11, 2002, from http://www.pmi.org/standards/wbspractice.htm

Colkin, E. (2002, October 21). Getting tough on ROI. *InformationWeek*. Retrieved October 23, 2002, from http://www.informationweek.com/story/IWK20021017S0013

General Services Administration (2002). *General Services Administration, fiscal year 2003 budget overview*. Washington, DC: U.S. Government Printing Office.

General Services Administration (2002). *Capital planning and investment control for IT*. Washington, DC: Government Printing Office. Retrieved May 8, 2002, from http://www.gsa.gov/Portal/content/pubs_content.jsp?contentOID=120784&contentType=1008

Hay, D. C. (2000). *A different kind of life cycle: The Zachman framework*. Retrieved May 15, 2002, from http://www.essentialstrategies.com/publications/methodology/zachman.htm

Mello, A. (2001, October 4). Why ROI can sometimes lie. ZDNet. Retrieved May 8, 2002, from http://techupdate.zdnet.com/techupdate/stories/main/0,14179,2816181,00.html

Netcraft.com. *Netcraft web server survey*. Retrieved May 22, 2002, from http://www.netcraft.com/survey/

Office of the Secretary for Defense (2000). *Contract pricing reference guides, vol 2. Quantitative techniques for contract pricing*. Retrieved November 16, 2002, from http://www.acq.osd.mil/dp/cpf/pgv1_0/pgv2/pgv2c9.html

Porter, R., & Lees, J. (2002). *Metrics improve software cost estimating, General Dynamics electronic systems*. Retrieved August 23, 2002, from http://www.galorath.com/client_case-dynamics.shtm

Remenyi, D., Money, A., Sherwood-Smith, M., & Irani, Z. (2000). "Why evaluation information technology investments?" In Butterworth & Heinemann (Eds.), *The effective measurement of management of IT costs and* benefits (2nd ed.) (pp. 33–35). Woburn, MA: Butterworth Heinemann.

Shulte, R. (2001). *What is the health of my project: The use and benefits of earned value.* Retrieved August 16, 2002, from http://www.allpm.com/article.php?sid=224

Stein, T. (1999, May 24). Making ERP add up. *InformationWeek*. Retrieved May 28, 2002, from http://www.informationweek.com/735/prerp.htm

Surmacz, J. (2002, April 10). Nuts-and-bolts investment. *CIO Magazine*. Retrieved May 8, 2002, from http://www2.cio.com/metrics/2002/metric349.html

University of Southern California, Center for Software Engineering (2002, September 23). *COCOMO II*. Retrieved October 23, 2002, from http://sunset.usc.edu/research/COCOMOII/cocomo_main.html

US Department of Labor, Bureau of Labor Statistics (2001). *Industry labor productivity tables, 1987 forward, all published 4-digit industries*. Retrieved May 21, 2002 from http://www.bls.gov/lpc/iprdata1.htm#Industry%20Labor%20Productivity%20Tables,%201987%20Forward,%20All%20Published%204-Digit%20Industries

Wu, J. (2000, December 8). *Calculating ROI for business intelligence projects part 1*. Retrieved May 22, 2002, from http://www.datawarehouse.com/iknowledge/articles/article.cfm?ContentID=936

FURTHER READING

Brooks, F. P., Jr. (1995). *The mythical man-month: Essays on software engineering, anniversary edition* (2nd ed.). Reading, MA: Addison–Wesley.

Guidelines and discount rates for benefit-cost analysis of federal programs (1992, October 29). Circular No. A-94, Revised (Transmittal Memo No. 64). Office of Management and Budget Web site/Bureau of Printing. Retrieved May 25 2002, from http://www.whitehouse.gov/omb/circulars/a094/a094.html#5

Colley, J. L., Jr., Doyle, J. L., & Hardie, R. D. (2002). *Corporate strategy*. New York: McGraw–Hill.

Davenport, T. H. (1999). "Putting the enterprise into the enterprise system." In *On the business value of IT, Harvard Business Review.* Cambridge, MA: Harvard Business School Press.

Dietel, H. M., Dietel, P. J., & Steinbuhler, K. (2000), *E-business and e-commerce for managers.* Upper Saddle River, NJ: Prentice–Hall.

Fellenstein, C., & Wood, R. (2000). *Exploring e-commerce, global e-business, and e-societies,*. Upper Saddle River, NJ: Prentice–Hall.

Financial Accounting Standards Board (2002). Retrieved November 2002 from Technical Inquiry Service, FASB: http://www.fasb.org/inquiry/

Harmon, P., Rosen, M., & Guttman, M. (2001). *Developing e-business systems and architectures: A manager's guide.* San Diego, CA: Morgan Kaufmann.

Investorwords.com (2002). The Internet Society, statistics page. Retrieved August 23, 2002, from: http://www.isoc.org/internet/stats/

IT infrastructure outsourcing (2002). Retrieved October 2002 from http://www.csc.com/solutions/itinfrastructureoutsourcing/

Kalakota, R., & Robinson, M. (2001). *e-Business 2.0: Roadmap for success*. Reading, MA: Addison–Wesley.

Measuring the value of information technology (2001). Presented at the E-Investments for CFOs Conference, New York, October 1–2, 2001. Available at iValue.com: http://www.ivalueinstitute.com/

Parker, M. M., Benson, R. J., & Trainor, H. E. (1988). *Informatione economics; Linking business performance and information technology* (10th ed.). Upper Saddle River, NJ: Prentice–Hall.

Remenyi, D., Money, A., Sherwood-Smith, M., & Irani, Z. (2000). Why evaluation information technology investments? In *The effective measurement of management of IT costs and benefits* (2nd ed.). Woburn, MA: Butterworth–Heinemann.

Ryssel, R., Ritter, T., & Gemunden, H. G. (2000). The impact of IT on trust, commitment and value-creation in enter-organizational customer-supplier relationships. Retrieved November 2002 from http://www.bath.ac.uk/imp/pdf/114_RysselRitterGemuenden.pdf

Shaker, S. M., & Gembecki, P. M. (1998). *The war room: Guide to competitive entelligence*. New York: McGraw-Hill.

Technology outsourcing (2002). Retrieved October, 2002, from http://www-1.ibm.com/technology/offerings/outsourcing.shtml

The BroadVision products page (2002). Retrieved September 2002 from http://www.broadvision.com/OneToOne/SessionMgr/products/products_main.jsp

The Internet Society (2002). *Statistics page*. Retrieved May, 2002, from http://www.isoc.org/internet/stats/

WebSphere software platform (2002). Retrieved October 2002, from http://www-3.ibm.com/software/info1/websphere/index.jsp

E-government

Shannon Schelin, *North Carolina State University*
G. David Garson, *North Carolina State University*

INTRODUCTION

Information technology has fundamentally altered the way we interact in today's society. The advent of personal computers, information networks, and the Internet has engendered an information revolution. E-government has developed from the information revolution as a new method of connecting with all constituents of government. Essentially, e-government consists of new channels of communication and new methods for participation between governments, citizens, employees, and businesses. Although there is widespread interest in the topic, e-government lacks a common definition. The American Society for Public Administration (ASPA) and United Nations Division for Public Economics and Public Administration (UNDPEPA) have defined e-government as "utilizing the Internet and the World Wide Web for delivering government information and services to citizens" (UN & ASPA, 2001, p. 1). Furthermore, it is often summarized as a process revolutionizing the business of government through the use of information technology (IT), particularly Web-based technologies, which improve internal and external processes, efficiencies, and service delivery.

What E-government Is and How It Is Used

E-government has evolved from the information technology revolution, particularly the proliferation of the World Wide Web. IT in government has long been acknowledged as a method for improving efficiency and communication (Kraemer & King, 1977; Norris & Kraemer, 1996). The advent of the Internet has led to developments such as electronic mail (e-mail), which have profound organizational consequences through the perceived erosion of limitations such as time and geography (Rahm, 1999). The main concerns of e-government move beyond the benefits of automating processes and employing new public ITs and concentrate more on reinventing processes to exploit fully the potential of information technology. E-government involves government to citizen (G2C), government to government (G2G), government to business (G2B), and government to employee (G2E) interactions.

According to some proponents of e-government, it is more than just a shift in communication patterns or mediums. Theoretically, e-government involves a transformation of the organizational culture of the government. According to Fountain (2002), the cultural transformation that can be achieved through e-government resides in the elimination of redundancy and duplicity of service provision. It should be noted, however, that these shifts in culture are not without difficulty. To reorient government culture successfully, data must be shared both vertically and horizontally. By creating a distributed data environment, the system will create winners and losers in terms of control over data. This battle, both political and bureaucratic, will hinder the cultural shifts associated with the apogee of e-government.

Despite the issues associated with cultural reorientation, recent authors argue that information technology facilitates new governmental structures and parameters as demanded by citizens and businesses (Heeks, 1999; Ho, 2002; Osborne & Gaebler, 1992). These demands require services that cut across traditional departmental and agency lines, which in turn require government to improve communication and interaction (Alexander & Grubbs, 1998). These new requirements, which fundamentally alter the nature of government, are made possible through the strategic use of information technology. As Fountain (2001) noted, the reinvention process requires overcoming the rigidities and limits of traditional bureaucracy. Specific objectives may include the centralization of public data and the improvement of internal processes and communications (Alexander & Grubbs, 1998). This delivery of services and information also involves the integration of government networks and databases to allow for cross-agency communication and interaction (Moon, 2002). Again, the turf issues associated with the cultural context of government may preclude some of these efforts from reaching maturation. The impact of the information era in government can be witnessed in the increasing governmental movement to Web-based information, services, and transactions, however, as evidenced by recent national surveys (International City/County Management Association [ICMA], 2002; West, 2000). Prior to looking at specific data about the expansion of e-government in the United States, it is important to examine the history of federal legislation regarding public information technology and e-government.

History of E-government at the U. S. Federal Level

The U.S. groundwork for e-government at the federal level was laid in the 1980s. The Paperwork Reduction Act of 1980 mandated an information resources management (IRM) approach to federal data, thereby establishing for the first time a single policy framework for information resource management at the federal level. The director of the Office of Management and Budget (OMB) was given responsibility for developing an IRM policy and for overseeing its implementation. A major revision of the Paperwork Reduction Act in 1995 mandated strategic planning for IRM.

At the same time this administrative structure was created, the physical Internet was being born. ARPANET, a project of the Department of Defense, had gone online in 1969, uniting some 23 defense-related military, university, and research laboratory sites by 1973. It was the funding of NSFNet by the National Science Foundation in 1986, however, that really marked the beginnings of the modern Internet. This long-haul backbone network was placed under a cooperative agreement with IBM, MCI, and Merit Network the following year and by the end of 1987 there were 10,000 Internet hosts. By 1989, there were 100,000. By the time Tim Berners-Lee developed the World Wide Web in 1992 and Mosaic software was released to surf the Web, the number of Internet hosts exceeded 1 million. One of these was the server that housed the first White House home page, launched on the Web in 1992.

Up to 1992, access to the Internet backbone was limited by the NSF's Acceptable Use Policy, which prohibited commercial traffic on the Internet. With the support of Congressman Rick Boucher (D-VA), chairman of the Science Subcommittee of the House Committee on Science, Space, and Technology, legislation was passed, and in November 1992, President Bush signed new legislation that repealed the Acceptable Use Policy, replacing it with language that permitted commercial traffic on the Internet backbone. In 1993, federal funding of the Internet ended and the Internet became a private sector entity.

With the full advent of the "Internet Age" in the early 1990s, early efforts that might be seen as part of building e-government centered on educational functions. The Information and Technology Act of 1992 sought to ensure technology development in public education, health care, and industry. It called on NSF to fund efforts to connect K–12 classrooms to the Internet. In 1993, the National Information Infrastructure Act was passed, targeting federal research and development funding to the accelerated development of high-performance computing and high-speed networking services at the university and research laboratory level.

An important part of the administrative background of e-government in the early to mid-1990s was the National Performance Review created March 3, 1993, under the leadership of Vice President Al Gore, himself a strong technology advocate. The National Performance Review reflected the Clinton administration's emphasis on IT as a tool to reform government. Its report, *Creating a Government that Works Better and Costs Less: Reengineering Through Information Technology,* was an important document of the "reinventing government" movement. This administrative movement originated with a focus on traditional forms of decentralization, devolution, and privatization. "Reinvention" advocates quickly came to see, however, that e-government was a major reform thrust facilitating and implementing their own goals of a results-oriented, client-centered, market-reliant approach to government. In 1993 the Government Information Technology Services Board was created to help implement NPR in information technology areas.

As the momentum for e-government grew, so did awareness of a major impediment: the skew of access to the Internet by the well-to-do and lack of access by America's "have-nots." The Commerce Department's 1994 report, *Falling Through the Net,* brought public attention to the issue of the "digital divide." In response, the Telecommunications Act of 1996 provided for a Universal Service Fund fee (a telephone tax, also known as the "E-rate" fund or fee), part of which was used by the Clinton administration to provide modem-based Internet access to schools, libraries, Indian reservations, and other "digital divide" target groups. By the end of the decade, however, the digital divide problem had receded because of falling hardware and software prices, making Internet access affordable to the great majority of citizens.

Some credit must also be given to another Clinton administration initiative—openness in government. The Electronic Freedom of Information Act Amendment of 1996 extended the right of citizens to view executive agency records to include access to electronic formats and online opportunities for access information.

Also in 1996 came a piece of legislation that provided the basis for centralizing in the Office of Management and Budget the authority to enforce implementation of e-government (as well as other IT practices) across the federal government. The Clinger–Cohen Act of 1996 (originally named the Information Technology Management Reform Act of 1996, an amendment to the Paperwork Reduction Act of 1980) established a chief information officer in every agency, making agencies responsible for developing an IT plan which relates IT planning with agency missions and goals. The oversight role of the director of the OMB was strengthened and later, when e-government became a priority, the existence of the chief information officer strategic planning structure was an important element facilitating e-government implementation at the federal level.

A variety of agency-centric developments extended e-government in the late 1990s. In 1997, the U.S. Department of Agriculture became the first agency to engage in e-rulemaking, soliciting Web-based comments on rules for organic foods. This initiative won the 1998 Government Technology Leadership Award and became the basis for one of the major projects of the e-government efforts of the current Bush administration (see the President's E-Government Strategy, available at http://www.whitehouse.gov/omb/inforeg/egovstrategy.pdf). The 1998 amendments to the Rehabilitation Act required federal agencies to make their IT and electronic information available to people with disabilities. The Internal Revenue Service (IRS) Restructuring and Reform Act of 1998 promoted electronic filing of tax returns, requiring the IRS to establish that all forms, instructions, publications, and other guidance are available via the Internet. It also provided for taxpayer electronic access to their accounts by 2006.

In some ways, the official "start" of federal e-government might be traced to the Government Paperwork Elimination Act of 1998, which authorized the OMB to acquire alternative ITs for use by executive agencies; support for electronic signatures; and electronic filing of employment forms, with a deadline set for most forms to be in place by October 21, 2003. The Government Paperwork Elimination Act was the legal framework for accepting electronic records and electronic signatures as legally valid and enforceable and also represented Congressional endorsement of the e-government strategy (Fletcher, 2002).

In a Presidential Memo of December 17, 1999, titled *Electronic Government,* President Clinton endorsed the concept of a federal governmentwide portal (later to be FirstGov.gov). In 2000, the President's Management Council adopted digital government as one of its top three priorities. On June 24, 2000, President Clinton made the first presidential Internet address to the nation, again calling for establishment of the FirstGov.gov portal. FirstGov.gov was launched September 22, 2000 as a Clinton management initiative. It is the official U.S. government portal, designed to be a trusted one-stop gateway to federal services for citizens, businesses, and agencies. At launch, it was a gateway to 47 million federal government Web pages. FirstGov.gov also links to state, local, District of Columbia, and tribal government pages in an attempt to provide integrated service information in particular areas, such as travel. The Office of Citizen Services and Communications, within the General Services Administration, manages FirstGov.

In Election 2000, both candidates (Gore and Bush) advocated digital government concepts. After the election, President Bush made e-government one of his central management reform themes (although opposing most "digital divide" funding advocated by Clinton and Gore). *The President's Management Agenda,* issued in August 2001, committed the Bush Administration to five major management objectives, one of which was electronic government.

In June 2001, the OMB created the position of associate director for information technology and e-government. This gave the OMB a key point of contact to give higher priority to IT initiatives, particularly the goal of creating a citizen-centric government. In essence, this position had a mandate to provide leadership to all federal IT implementation, including a special emphasis on e-government. The first incumbent was Mark Forman, who quickly took strong actions to implement e-government throughout federal agencies. On July 25, 2002, the first chief technology officer for the federal government was appointed, responsible for overseeing implementation of e-government policies. The first incumbent was Casey Coleman, heading up the General Services Administration's Office of Citizen Affairs.

The OMB issued the cornerstone document, *E-Government Strategy,* on February 17, 2002. This document set forth Bush administration e-government principles: citizen-centric, results-oriented, and market-based. It also called for increased cross-agency data sharing. Some 34 specific projects were identified for funding, including those in the "Quicksilver Initiative" announced in October 2001, as noted in Table 1.

Table 1 Overview of U.S. Federal E-government Projects

Government to Citizen (G2C)
USA Service (GSA)
EZ Tax Filing (Treasury)
Online Access for Loans (DoEd)
Recreation One Stop (Interior)
Eligibility Assistance Online (Labor)
Government to Business (G2B)
Federal Asset Sales (GSA)
Online Rulemaking Management (DOT)
Simplified and Unified Tax and Wage Reporting (Treasury)
Consolidated Health Information (HHS)
Business Compliance One Stop (SBA)
International Trade Process Streamlining (Commerce)
Government to Government (G2G)
E-Vital (SSA)
E-Grants (HHS)
Disaster Assistance and Crisis Response (FEMA)
Geospatial One Stop (Interior)
Wireless Networks (Justice)
Internal Effectiveness/Efficiency (Government to Employee (G2E))
E-Training (OPM)
Recruitment One Stop (OPM)
Enterprise HR Integration (OPM)
Integrated Acquisition (GSA)
E-Records Management (NARA)
Enterprise Case Management (Justice)

Among other recent developments in e-government at the federal level has been the implementation of e-procurement. The General Services Administration and Office of Federal Procurement Policy, with involvement from the Department of Defense, the National Aeronautics and Space Administration, and National Institutes of Health, advanced e-procurement by establishing Past Performance Information Retrieval System to give online access to past vendor performance records in 2002. Another development is the Dot Kids Implementation and Efficiency Act of 2002, passed in December 2002. This act created a new domain, like.com and like.edu. Every site designated ".kids" will be a safe zone for children and will be monitored for content, for safety, and all objectionable material will be removed.(Many civil liberties proponents have criticized this act as allowing government censorship thereby reducing transparency.)

The Electronic Government Act of 2002 was passed by Congress November 15, 2002, and signed by the president on December 16, 2002. The act was sponsored by Senator Joseph Lieberman (D-CT). It is intended to promote e-government in all federal agencies. The Electronic Government Act establishes an Office of Electronic Government within the Office of Management and Budget. The head of this office is to be appointed by the president and report to the OMB director. In essence, this formalizes the administrative setup established by the OMB in 2001 under Mark Forman. It requires regulatory agencies to publish all proposed rules on the Internet and to accept public comments via e-mail as part of "e-rulemaking." All information published in the Federal Register must now be published on the Web also. The federal courts must publish rulings and other information on the Web, and there are numerous other provisions. The Electronic Government Act makes $45 million available for e-government projects in the current fiscal year 2003 and $345 million over five years. (This compares with $5 million for Forman in fiscal year 2002).

Clearly, the legislation and mandates of the U.S. federal government indicate the importance of e-government. Other countries have even more sophisticated laws and procedures to guide e-government adoption. For example, Singapore has developed long-range strategic plans for e-government, incorporating the eCitizen Center as the one-stop shop for all G2C interactions. The United Kingdom also demonstrates one of the highest levels of executive commitment to e-government. The Office of the E-Envoy has been created to oversee the transition to e-government for all services by 2005. Many other examples from around the globe further demonstrate the shift toward e-government.

E-GOVERNMENT ADOPTION IN THE UNITED STATES

Following the detailed review of U.S. federal commitment to e-government, it is important to assess the levels of e-government adoption at the local government level, as well to illustrate usage by citizens. This description focuses on the United States; however, other sources provide more international coverage (Demchak, Friis, & La Porte, 2001; UN & ASPA, 2002). By analyzing data obtained from the 2002 E-Government Survey conducted by the International City/County Management Association, the adoption rates of e-government at the local level can be assessed. The survey was conducted to evaluate the involvement of local governments in e-government activities, including Web site development, electronic services, geographic information systems, changes associated with e-government adoption, and barriers preventing such adoption. The survey was sent to 7,844 municipal and county governments with populations over 2,500; 4,123 surveys were completed and returned, a response rate of 52.6%.

For the purposes of this analysis, the municipalities and counties have been divided into three population categories. Small jurisdictions contain less than 50,000 inhabitants, medium jurisdictions contain 50,000 to 249,999 inhabitants, and large jurisdictions contain 250,000 or more. The responding jurisdictions represent all four geographic regions: Northeast, North-Central, South, and West, as defined by the ICMA.

The presence of local government Web sites, as gathered during the ICMA survey, is significant, with 74.2% of responding counties and municipalities having an official Web site. In terms of the municipalities, 71.1% of small jurisdictions, 97.9% of medium jurisdictions, and 100% of large jurisdictions have official Web sites. Furthermore, 58.8% of small counties, 92.8% of medium counties, and 98.4% of large counties have an official Web presence. The high rate of Web presence demonstrates the movement of local governments to the e-government model; however, it is important to note that static Web presence is not sufficient for true e-government.

To better assess the required interactive component of e-government, the presence of Web-enabled transactional processes via governmental Web sites is analyzed. The ICMA survey queried respondents about several categories of transactions, including online payments of taxes, utility bills, fines/fees, form completion and submission, business license application/renewal completion and submission, online requests for records, online delivery of records, online requests for services, online registration for facility usage, online voter registration, online property registration, downloadable forms, and online communication with elected and appointed government officials. In the aggregate, few local governments are using Web-based transactions. Table 2 demonstrates the offering of transactions by municipalities, based on the population categories previously denoted.

The positive effect of size on the rate of transaction offerings is important to note. For example, only 1.8% of small municipalities offer online payment of taxes, whereas 5.3% of medium and 24.0% of large municipalities offer the same online service.

A similar trend regarding the increased offering of Web-based transactions based on population size is found in the counties. Table 3 highlights the percentage of counties offering specific transactions, by population grouping.

Although electronic transaction offerings are limited in local governments, there is an increasing citizen and business demand for such Web-based services. A national survey conducted by the Pew Research Center during September 2001 demonstrates the widespread usage of governmental Web sites and their interactive services. The 2002 Pew Internet and American Life Project indicates that 58% (68 million people) of American Internet users have accessed at least one governmental Web site (Larson & Rainie, 2002). Furthermore, 68% of Internet users indicate that government Web sites improve their interaction with at least one level of government. Approximately 63% of those surveyed have downloaded government forms, 16% have filed taxes online, and 12% have renewed automobile registrations online. The 2002 Pew Internet and American Life Project highlights the citizen interest in and use of the Internet to enhance standard interactions with government.

Although the use of the Internet by governments is increasing, there is still a significant lag time compared with private sector adoption of new technologies. Both structural inertia and risk aversion, commonplace in the public sector, foster governments that are slow to adopt and implement new technologies. The inherent tension between the need for reliability and accountability contrasted with reliance on maintaining organizational status quo leads to the increased adoption lag time in governmental organizations. For government IT adoption and implementation, the inertia that exists within the public sector means that organizations are often less willing and able to engage

Table 2 Percentage of Transactional Offerings by Municipalities, by Population Grouping

SERVICE	SMALL MUNICIPALITIES	MEDIUM MUNICIPALITIES	LARGE MUNICIPALITIES
Online payment of taxes	1.8	5.3	24.0
Online payment of utility bills	3.0	13.5	15.4
Online payment of fines and fees	2.1	12.4	23.1
Form completion and submission	7.7	16.1	33.3
Business license application and renewal completion and submission	4.4	8.0	14.8
Online requests for records	26.4	31.9	38.5
Online delivery of records	16.2	21.4	25.9
Online requests for services	31.3	41.0	44.4
Online registration for facility usage	12.1	20.3	36.0
Online voter registration	1.5	1.7	9.5
Online property registration	2.1	5.0	5.3
Downloadable forms	50.2	75.3	92.0
Online communication with elected and appointed officials	68.7	84.2	96.0

new technologies (Bretschneider, 1990). This inability to adapt to the rapidity of the information age creates a disconnect between the movement toward e-government and the need to consider all citizen requirements and needs. The most common example of this problem is demonstrated by the creation of governmental Web sites that are organized by administrative structure rather than by service orientation, which is more intuitive to constituents. To understand more fully the issues associated with e-government adoption, including challenges and benefits, it is crucial to understand the theory and typology associated with e-government.

THEORY AND TYPOLOGY OF E-GOVERNMENT

Using the theoretical frameworks of e-government of decentralization–democratization and normative–dystopian, as outlined by Garson (1999), one can begin to assess the benefits and challenges associated with e-government. The decentralization–democratization framework of e-government revolves around the progressive nature of technology and highlights governmental advances resulting from e-government. Essentially, decentralization–democratization highlights the fundamental restructuring of government interactions with its constituents through the strategic use of information technology, that is, greater efficiency, effectiveness, and increased participation. On the other hand, the normative–dystopian framework uses the high rates of conflict and failure associated with information technology applications and offers a pragmatic, even skeptical view of e-government. The normative–dystopian framework is associated with the notion that e-government is another medium for communication between constituents and the government; however, it does not support the concepts of reinvention

Table 3 Percentage of Transactional Offerings by Counties, by Population Grouping

SERVICE	SMALL COUNTIES (%)	MEDIUM COUNTIES (%)	LARGE COUNTIES (%)
Online payment of taxes	9.2	25.0	40.0
Online payment of utility bills	1.0	8.2	9.4
Online payment of fines and fees	1.9	9.2	11.1
Form completion and submission	5.6	13.6	21.1
Business license application and renewal completion and submission	1.9	5.2	16.1
Online requests for records	23.1	43.2	65.0
Online delivery of records	12.0	35.9	53.6
Online requests for services	7.6	17.9	25.9
Online registration for facility usage	6.7	10.6	21.4
Online voter registration	4.7	7.0	7.4
Online property registration	1.2	2.7	8.3
Downloadable forms	42.1	72.6	87.7
Online communication with elected and appointed officials	63.6	77.1	82.1

of government often associated with e-government. Although neither framework can be considered fully descriptive, taken together they provide a useful delineation of the theoretical literature on e-government.

Decentralization–democratization theory is the traditional view of e-government, beginning with Bozeman and Bretschneider's seminal article in 1986, which highlights the transformational, progressive nature of technology adoption in the government sector. Furthermore, Reschenthaler and Thompson (1996) contended that the power of public information technology, a prerequisite for e-government, lies in its ability to level the playing field for all sizes and types of governments. They see the basis for reengineering the business of government, refocusing its work on the needs of the citizens, and returning government to its core functions residing under the technology umbrella.

Another decentralization–democratization approach for examining the e-government model involves comparing and contrasting traditional bureaucratic design with the evolved e-government design. The traditional bureaucratic model of public service delivery (the Weberian model) focuses on specialization, departmentalization, and standardization (Ho, 2002). This traditional model has created departmental "silos" that resist functioning across agency boundaries, in the name of equitable and efficient governmental interactions. In the 1990s, however, the reinventing government movement sought to shift the core focus of government, moving from departmentalization and centralization to citizen-centric decentralization (Osborne & Gaebler, 1992).

The e-government paradigm, which emphasizes coordinated network building, external collaboration, and customer services, is slowly replacing the traditional bureaucratic paradigm and its focus on standardization, hierarchy, departmentalization, and operational cost-efficiency (Ho, 2002). In this view, the development and deployment of technology networks, shared databases, and Web-based interactions are the necessities that will facilitate seamless integration of government services. Although this shift to the e-government paradigm has not been fully realized, the Internet is a key enabler to this transformation because it provides government with the ability to use technology to impact customers directly, instead of simply reengineering internal processes (Scavo & Shi, 1999).

Decentralization–democratization proponents uses technology and e-government adoption rates as proxies for gauging the success of e-government. As the adoption and implementation rates increase, e-government technologies will enjoy increased legitimacy (Fletcher, 1999). Increasing citizen and business demand for e-government applications, which have permeated to the local government level, are central to greater adoption across all levels of government (Norris, Fletcher, & Holden, 2001). Response to the changing environment with improved service delivery, increased efficiency, and reduced costs is key to e-government success (West, 2000). Beyond the optimistic prospects of the democratization/decentralization theorists, there is a concrete reality of governmental technology and e-government failures, which are explained in the normative–dystopian framework.

Issues of privacy, security, and the digital divide fall under normative–dystopian model, which offers a critical approach to evaluating e-government. This view entertains concepts of dehumanization and isolation, resulting from the proliferation of information technology. The normative–dystopian theory addresses ethical issues that surround e-government. Recent concerns about the "digital divide," the technology gap that exists between distinct groups in the United States, highlight the issues associated with the move to e-government as a mode of service delivery, due to the potential consequences for unconnected or underserved populations. Several studies note racial, regional, educational, gender, and age disparities among Internet users and technology owners (Norris, 2001; Novotny, 1998). These gaps are of great concern for public administrators, who must serve efficiently, effectively, and equitably to fulfill their public charge.

Issues of security and privacy in the post–September 11 era are of grave concern to the majority of government constituents. In fact, the 2002 ICMA E-Government Survey indicates that 15% of the survey respondents have or will make significant changes to their existing security practices as a result of the terrorist attacks. Additionally, more than 10% indicate that they have removed information from their Web site for security reasons. Clearly, the issues of security and privacy are more salient than ever and governments must address constituent concerns if e-government is to achieve high levels of adoption and usage.

Using the normative–dystopian framework, Bovens and Zouridis (2002) examine the inherent problems associated with the shift toward an e-government paradigm. The emerging emphasis on information technologies as the medium for citizen interaction with government fundamentally alters the role of the bureaucrat. The traditional Weberian model uses street-level bureaucrats to interact with citizens and to determine the proper services and service levels to assist these citizens (Lipsky, 1980). This method allows for expertise, judgment, and practicality to be engaged in the decision-making process. In the e-government model, computer programs are used to interface with clients, assess eligibilities, and decide on proper levels of service (Boven & Zouridis, 2002). This model suggests that the street-level bureaucrats are losing their discretionary power, which can have deleterious effects on the clients. Bovens and Zouridis use the normative–dystopian framework to highlight the potential arbitrariness and threats to the legitimacy of governmental actions at the street level in the e-government model. Both frameworks for assessing e-government are valuable and valid; however, the rising citizen demand and increasing governmental use of e-government components indicate that e-government will only progress in future years.

PHASES OF E-GOVERNMENT

Several models have been developed to explain the progression of e-government (Layne & Lee, 2001; Moon, 2002). The stages or iterations of e-government do not follow a linear progression, however. Often, e-government applications are developed for utility, an ad hoc approach, instead of according to a master project plan, a systematic approach. For example, one local government may offer the ability to pay property taxes online through an application service provider but still have a basic informational Web site without interactive enhancements. Furthermore, the various models to be described offer a broad categorization of each stage or iteration. These are arranged in a continuum in which governments can be within the same stage while having diverse service capacities or functionalities.

According to the UN and ASPA (2001), there are five main stages of e-government. These five stages are precipitated by the recognition that some form of presence exists for the given locality. Stage One, the emerging Web presence, presents static information and is often considered to be "brochure-ware" (UN & ASPA, 2001, p. 16). The main goal of this stage is to provide an online dissemination of general information about the government. Often, the Stage 1 Web site visually represents the "stovepipes" that exist within agencies and does not allow for continuity across various departmental pages.

The association of the Web site denotes enhanced Web presence, Stage 2, with information on services; however, it is still organized by departments. This stage often offers e-mail as a method of two-way communication. It offers limited communication and greater information about the services of the government but is not consistent with the citizen-centric approach that has been advocated for e-government (UN & ASPA, 2001, p. 17; Layne & Lee, 2001; Moon, 2002).

Stage 3, interactive Web presence, does offer some of the citizen-centric methods as advocated by e-government proponents (UN & ASPA, 2001, p. 18). Information is presented in intuitive groupings, rather than by departmental or agency association. Often, portals are used as a single point of entry into various departments and service areas. Major groupings in the portal design might include business, new resident, seniors, children, or other standard groups. Subsequently, end users would control their Web destination based on the grouping they select. Each sublevel under the group headings offers commonly requested services, information, and assorted items of interest to the particular group. Again, specific agency or department does not designate the services and information contained within each group; they are offered as a bundle of interest to the target population. Stage 3 sites also have downloadable forms with online submissions, e-mail contact for various governmental employees, and links to other governmental Web sites.

Stage 4, transactional Web presence, allows secure online transactions (UN & ASPA, 2001, p. 19). User needs dominate the organization of this Web presence and the information presented is dynamic. Potential transaction offerings include online payments of taxes, utility bills, and fines and fees; form completion and submission; business license application and renewal completion and submission; online registration for facility usage; online voter registration; and online property registration. As evidenced by the ICMA 2002 E-Government Survey, less than 15% of responding counties and municipalities offer these services. Furthermore, only 4.6% of the respondents offer online payments.

The final stage, Stage 5, involves seamless government. This stage represents an ideal and no real example of its application is available. Stage 5 ideally involves a cross-agency, intergovernmental approach, however, that only displays one front, regardless of service area (UN & ASPA, 2001, p. 20). For example, a portal may offer a compendium of local, state, and federal government services without user recognition of what level of government provides the service. A Stage 5 site would offer vertical and horizontal integration and would require true organizational transformation with respect to administrative boundaries (UN & ASPA, 2001, p. 20).

Drawing on the UN and ASPA typology, current literature indicates that the majority of local governments are in Stage 2, with enhanced Web presence. Moon's (2002) analysis of 2000 ICMA E-Government Survey indicates that a majority of municipalities with populations over 10,000 are not offering transactional Web sites. Furthermore, based on the 2002 ICMA E-Government Survey, only 62 (1.7%) municipalities offer online payment of taxes, 113 (3.1%) offer online payment of utility bills, and 91 (2.5%) offer online payment of fines and fees. The percentages for counties are only slightly higher, with 69 (16.3%) offering online payment of taxes, 17 (4.0%) offering online payment of utility bills, and 21 (5.0%) offering online payment of fines and fees. More robust transactional services can be found at the state and federal levels of government, as would be expected.

Using Ho's (2002) methodology for assessing municipal Web sites, a different dimension of e-government typology can be discerned. He centered on three primary orientations designed to demonstrate the shift from the Weberian bureaucratic model to the e-government paradigm. The first orientation is administrative-oriented Web sites, organized along departmental lines and represents the traditional bureaucratic paradigm (p. 437). The second orientation is information-oriented, crossing departmental lines to provide a one-stop shopping experience to the user (p. 437). The final orientation is user-oriented, which categorizes information in intuitive groupings that offer end user control and rapid access to predetermined information (p. 437). This orientation also crosses traditional departmental lines in an attempt to provide an all-encompassing experience to the end user. Ho's analysis of the 55 largest municipalities' Web sites indicates that a majority has moved toward varying degrees of user-orientation (p. 438).

Further alternative approaches to defining the e-government typology use various levels of communication, applications of technology, and citizen participation in democratic forums to define its stages. One example of this approach is found in Moon's typology (2002), as adopted from Hiller and Bélanger. This framework also uses five stages but its focus is on the communications between various stakeholders, including G2G, G2E, G2C. Stage 1 involves information dissemination and uses basic Web authoring tools and bulletin boards as methods of communication (p. 426). Two-way communication, using e-mail and electronic data exchange (EDI), is found in Stage 2. Stage 3 highlights service and financial transactions, along with technologies such as EDI, electronic filing systems, digital signature, and public key infrastructure. Stage 4 centers on the concepts of vertical and horizontal integration of the technologies found in Stages 1, 2, and 3. It is similar to Stage 5 of the UN and ASPA model in its seamless outward appearance. Political participation is the mainstay of Stage 5 in Moon's typology. It involves online voting, e-democracy, and e-participation. Very few local governments have evolved to Stage 5, evidenced by the fact that only 2.1% of respondents to the 2002 ICMA E-Government Survey are using online voter registration. This alternative approach to the typology of e-government still acknowledges the importance of the user's view of e-government but extends the model to include digital civic engagement and enhanced electronic democracy.

The various models of e-government allow for broad comparisons across the various governmental Web sites, using similar benchmarking criteria. Such comparisons are useful to the extent that they can highlight the critical success factors associated with individual experiences, which can be generalized to the general population. One of these critical success factors is training for e-government. Issues of training are directly related to the need to move along the e-government continuum because training fosters greater understanding of the processes and applications, as well as provides a method of eliciting user support. By examining the current training practices of governments with regard to e-government, one can begin to understand the need to bridge e-government expansion and training to create a holistic, enterprise approach.

TRAINING FOR E-GOVERNMENT

As the information age advances, citizens expect a more responsive and accountable government. The concept of "citizens online instead of in line" enables people to have immediate access and responsiveness, while reducing personal interactions with government employees. As e-government continues to expand at all levels of government, federal, state, and local governments need to prepare their elected officials and employees to handle the multitude of changes it entails. Citizen demand for and governmental usage of e-government, as evidenced by national surveys and Web assessments (Larsen & Rainie, 2002; West, 2000), indicate that such applications will only increase in usage and importance over time. This advances the need for assessing the current state of training for e-government to determine areas of current success as well as to ascertain future direction for training.

The majority of the literature concerning training for e-government involves analysis of collegiate public administration program curricula. As early as 1988, the National Association of Schools of Public Affairs and Administration (NASPAA) added computing as a skill set for accredited master of public administration (MPA) programs (Northrop, 2003). Further calls for the incorporation of IT education came from Perry and Kraemer (1993). In 1998, Brown and Brudney completed a comprehensive examination of 106 MPA programs and found that only about 30% of the schools included in the sample offered instruction on technology planning, policy development, and evaluation, despite the NASPAA recommendation to include these in the curriculum. By 2001, Kim and Layne had conducted an empirical study of student perceptions of digital government within their institution and set forth recommendations for future training, both in the schools of public administration as well as for the leaders of the public sector.

In an attempt to highlight the importance of high-quality training opportunities for federal, state, and local government elected and appointed officials, a data set from the 2002 E-Government Survey conducted by the ICMA was analyzed. Issues of staffing and training in local government units (counties and municipalities), as well as implementation styles and challenges, were selected to measure the need for high-quality e-government training. As noted previously, this data set provides a snap shot of the current status of information technology in municipalities and counties with populations over 2500.

Although e-government adoption is the first step toward an electronically enabled government, the manner of application implementation is critical to understanding the nature of public sector information technology. As noted by various sources, staffing issues are one of the key factors in determining the success or failure of technology applications (Brown & Brudney, 1998). The existence of distinct information technology departments can be used as a proxy for the commitment of resources to information technology initiatives, including e-government. There is a large disparity between the size groupings of the jurisdictions, however. Only 28% of small municipalities have IT departments, whereas 84% of medium municipalities and 85.7% of large have such departments. In counties, 44.1% of small counties have IT departments, compared with 86.4% of medium and 98.4% of large counties. Along with the development of the chief information officer, the density of information technology departments, particularly in larger governmental units, indicates the importance of high-quality technology staff and, subsequently, their training.

Another key issue associated with preparing for e-government involves addressing those issues designated as barriers by local governments. According to the data collected via the 2002 ICMA E-Government Survey, several barriers hinder local government movement to successful e-government implementation. The lack of technology/Web staff is the most commonly cited problem, across all types and sizes of local governments. More than 45% of small municipalities, 47.8% of medium municipalities, and 44.8% of large municipalities, indicate the lack of staff is their greatest barrier. Furthermore, 40.7, 46.1, and 22.6% of small, medium, and large counties, respectively, indicate the same concern. The issue of staffing can easily be resolved through proper training. One common solution found in local governments involves the promotion of a current staff member, often a town clerk or administrative assistant, to webmaster. This promotion can be an effective solution if the skill set required to undertake such as task is already in place or can be easily leveraged through training and education.

Another commonly cited barrier to successful e-government implementation is the lack of technology and Web expertise. More than 33% of small municipalities, 28.6% of medium municipalities, and 24.1% of large municipalities indicate that this is their greatest barrier.

It also is cited as a barrier by 29.9, 25.1, and 19.4% of small, medium, and large counties. Again, increasing the training opportunities available to government employees and officials can significantly reduce this barrier.

Beyond the barriers related to e-government implementation, it is important to consider the typical development and maintenance of e-government services. More than 60% of ICMA survey respondents offering e-government services indicate that they have e-government services developed in-house by staff members. Seventy-one% of small municipalities, 45.1% of medium municipalities, and 51.7% of large municipalities have in-house development of some e-government services. Additionally, 78.4% of small counties, 32.2% of medium counties, and 33.9% of large counties developed e-government services in-house. The decreasing reliance on in-house development of services as population size increases further demonstrates the need for high-quality training opportunities for all governments, especially critical for the smaller counties and municipalities who are relying on internal staff to design and develop their interfaces with the public.

Based on a brief survey of state governmental Web sites, there is evidence that most states offer type of training, although the offerings are typically limited to technology applications and software training. Very few sites offer comprehensive e-government training with the necessary organizational, political, and technological components. As outlined by Kim and Layne (2001), there are several core components to effective digital government training for public sector officials, including an understanding of e-government infrastructure; the political, economic, and legal contexts of e-government; best practices; conceptual issues such as portals and transactions; service-related issues pertaining to Internet-based offerings; and citizen-centric government (p. 238). Additionally, training should focus on the paradigmatic shift from the bureaucratic model of government to the e-government model. This component is based on Ho's (2002) premise that the bureaucratic paradigm, emphasizing standardization, hierarchy, departmentalization, and operational cost-efficiency, is slowly being replaced by the e-government model, emphasizing coordinated network building, external collaboration, and customer services.

Clearly, the advent of the information age has fundamentally changed daily life. Additionally, the role of the Internet continues to increase in importance as penetration and digital literacy expand. Citizen and business demands mandate that government become involved in digital government initiatives, while current successful efforts increase the legitimacy for further information technology adoption (Norris, 1999). Because of the growing critical mass related to e-government, federal, state, and local governments must prepare their elected officials and employees to manage and navigate the new landscape of the e-government model. There is a paucity of research dedicated to training for e-government, as well as a lack of training opportunities. These disparities must be redressed to provide a thoughtful, high-level knowledge base, which will engender successful e-government adoption and implementation.

CONCLUSION: THE FUTURE OF E-GOVERNMENT

The march of e-government has not been without setbacks, notably in the areas of universal access and governmental openness. Presidents Bush's 2003 budget proposed to eliminate two critically important community technology programs related to diminishing the "digital divide": the Technology Opportunities Program (TOP) and the Community Technology Center (CTC) initiative. A study by Brown University's Center for Public Policy studied 1,265 federal and state Web sites and found some 6% had restricted areas requiring a password to enter, an increase reflecting post–September 11 attention to security. Likewise, the Environmental Protection Agency has removed risk management plans and other information pertaining to hazardous waste sites from its Web site, citing risk of terrorism outweighing the right of citizens to know hazardous waste risks near their homes, workplaces, and schools.

Nonetheless, the setbacks are far overshadowed by the remarkably rapid expansion of e-government. The Pew Internet and American Life Project, in the Larsen and Rainie report cited earlier, stated, "The rise of e-government has been one of the most striking developments on the Web" (Larsen & Rainie, p. 5). Because of citizen demand for things such as e-access to building permits, dog licenses, and birth certificates, the bandwagon effect for implementing e-government is strong (Moulder, 2001). There is now considerable momentum for government change along the e-government model. Governments are competing to be seen as being on the leading edge, not laggard adopters (Sprecher, 2001).

Many cities are moving away from traditional bureaucratic emphasis on standardization, departmentalization, and operational cost-efficiency, toward the "e-government" paradigm, which emphasizes coordinated network building, external collaboration, and customer services (Ho, 2002). In this way the reinventing government movement is tied to the e-government movement.

Many municipal governments have adopted e-government, but it is still at an early stage and has not obtained many of expected outcomes (cost savings, downsizing, etc.) that the rhetoric of e-government has promised (Moon, 2002). (It should be noted that this "rhetoric of e-government" promises cost savings, etc., which arguably are not realistic outcomes of e-government.) There has been general progress in online services, privacy policy statements, and standardized navigational features but much less progress in disability and foreign language access. Most agencies used Web for one-way communication with no provision for two-way interaction with citizens (West, 2001). The desire to transcend limited progress is one motivation behind the push for public–private partnerships in support of e-government initiatives (Holmes, 2001).

It is still too early to assess e-government in terms of effect on democracy. Clearly there are great hopes that e-government will provide new opportunities for widening civic engagement and participation (Milward & Snyder, 1996). In principle e-government increases the reliability and accountability of public organizations. In

Santa Monica, CA, home to Public Electronic Network (PEN), one of the first major community computing experiments, city officials were eager to participate in a community computing experiment in its first year, but by the sixth year most had ceased participation, citing lack of substance of e-forums, too much “flaming” (searing e-mail messages in which the writer attacks another participant in malicious terms), and too many personal attacks by a minority of network participants (Docter & Dutton, 1998). The Electronic Government Act of 2002 has now mandated e-rulemaking, a critical aspect of e-democracy, and there are other e-democracy experiments (e.g., all hearings and committee sessions of the Michigan Legislature are now broadcast in streaming video, live over nine channels). Although e-government may have a much less profound effect on e-democracy than it is already having in the provision of governmental e-services, at the very least new channels of two-way communication between citizens and their government are being enabled.

GLOSSARY

E-democracy A range of reforms that use the Internet to increase public participation in governmental processes, including e-voting, e-rulemaking, and various forms of two-way interaction with elected and appointed government leaders using e-mail, e-forums, e-conferencing, and other methods.

E-government A strategy for making government more efficient and effective through e-services, e-procurement, and e-democracy.

E-procurement Purchasing goods and services online by governmental agencies, including use of the Internet to submit requests for proposals, requests for bids, acceptance of bids, and the letting and monitoring of performance contracts.

E-services The provision of government services via the Internet, including provision of online information, use of online forms of all types, the ability to make payments online, online educational services, online consulting services, and online screening and referral to services which cannot themselves be rendered online.

Portals The strategy of providing one-stop, cross-agency, and cross-jurisdictional gateways to government services in a given functional area, such as student needs, small business needs, or recreational needs, or based on an aggregation of user needs, such as all state services.

CROSS REFERENCES

See *Digital Divide; Global Issues; Internet Literacy; Nonprofit Organizations; Politics.*

REFERENCES

Alexander, J. H., & Grubbs, J. W. (1998). Wired government: Information technology, external public organizations, and cyberdemocracy. *Public Administration and Management: An Interactive Journal, 3.* Retrieved February 5, 2003, from http://www.pamij.com/

Bovens, M., & Zouridis, S. (2002). From street-level to system-level bureaucracies: How information and communication technology is transforming administrative discretion and constitutional control. *Public Administration Review, 62*(2): 174–185.

Bozeman, B., & S. Bretschneider. (1986). Public management information systems: Theory and prescription. *Public Administration Review, 46* [Special edition], 475–487.

Bretschneider, S. (1990). Management information systems in public and private organizations: An empirical test. *Public Administration Review, 50,* 536–545.

Brown, M. M., & Brudney, J. L. (1998). Public sector information technology initiatives. *Administration and Society, 30,* 421–443.

Demchak, C. C., Friis, C. C., & La Porte, T. M. (1999). Webbing governance: National differences in constructing the face of public organizations. In G. D. Garson (Ed.), *Handbook of public information systems* (pp. 179–196). New York: Marcel Dekker.

Docter, S., & Dutton, W. H. (1998). The First Amendment online: Santa Monica’s Public Electronic Network. In R. Tsagarousianou, D. Tambini, & C. Bryan (Eds.), *Cyberdemocracy: Technology, cities, and civic networks* (pp. 125–151). New York: Routledge.

Fountain, J. (2001). *Building the virtual state: Information technology and institutional change.* Washington, DC: Brookings Institution.

Fletcher, P. D. (1999). Strategic planning for information technology management in state governments. In G. D. Garson (Ed.), *Information technology and computer applications in public administration: Issues and trends* (pp. 81–97). Hershey, PA: Idea Group.

Fletcher, P. D. (2002). Government Paperwork Elimination Act: Operating instructions for an electronic government. *International Journal of Public Administration, 25,* 723–736.

Garson, G. D. (1999). Information systems, politics, and government: Leading theoretical perspectives. In G. D. Garson (Ed.), *Handbook of public information systems* (pp. 591–605). New York: Marcel Dekker.

Heeks, R. (1999). Reinventing government in the information age. In R. Heeks (Ed.), *Reinventing government in the information age* (pp. 9–21). New York: Routledge.

Ho, A. T.-K. (2002). Reinventing local government and the e-government initiative. *Public Administration Review, 62,* 434–444.

Holmes, D. (2001). *Egov: Ebusiness strategies for government.* London: Nicholas Brealey.

International City/County Management Association (ICMA) (2002). Electronic Government Survey, 2002. Retrieved February 5, 2003, from http://www1.icma.org/upload/bc/attach/{613FEECD-7345-40DF-971E-00E2E327A5AF}egov2002web.pdf.

Kim, S., & Layne, K. (2001). Making the connection: E-government and public administration education. *Journal of Public Affairs Education, 7,* 229–240.

Kraemer, K. L., & King, J. L. (Eds.). (1977). *Computers and local government.* New York: Praeger.

Larsen, E., & Rainie, L. (2002). The rise of the e-citizen: How people use government agencies’ Web sites. Pew Internet and American Life Project. Retrieved February 5, 2003, from http://www.pewinternet.org/reports/pdfs/PIP_Govt_Website_Rpt.pdf.

Layne, K., & Lee, J. (2001). Developing fully functional e-government: A four stage model. *Government Information Quarterly, 18,* 122–136.

Lipsky, M. (1980). *Street-level bureaucracy: Dilemmas of the individual in public services*. New York: Russell Sage Foundation.

Milward, H. B., & Snyder, L. O. (1996). Electronic government: Linking citizens to public organizations through technology. *Journal of Public Administration Research and Theory, 6,* 261–276.

Moon, M. J. (2002). The Evolution of E-government among municipalities: Rhetoric or reality? *Public Administration Review, 62,* 424–433.

Moulder, E. (2001). E-government—if you build it, will they come? *Public Management, 83,* 10–14.

Norris, D. F. (1999). Leading edge information technologies and their adoption: Lessons from US cities. In G. D. Garson (Ed.), *Information technology and computer applications in public administration: Issues and trends* (pp. 137–156). Hershey, PA: Idea Group.

Norris, D. F., Fletcher, P. D., & Holden, S. H. (2001). Is your local government plugged in? Highlights of the 2000 electronic government survey. Prepared for International City and County Managers Association and Public Technologies. Retrieved February 5, 2003, from http://icma.org/download/catIS/grp120/cgp224/E-Gov2000.pdf

Norris, D., & Kraemer, K. (1996). Mainframe and PC computing in American cities: Myths and realities. *Public Administration Review, 56,* 568–576.

Norris, P. (2001). *Digital divide: Civic engagement, information poverty, and the Internet worldwide*. Cambridge, England: Cambridge University Press.

Northrop, A. (2003). Information technology and public administration: The view from the profession. In G. David Garson (Ed.), *Public information technology.* Hershey, PA: Idea Group.

Novotny, P. (1998). The World Wide Web and multimedia in the 1996 presidential election. *Social Science Computer Review, 16,* 169–184.

Osborne, D., & Gaebler, T. (1992). *Reinventing government: How entrepreneurial spirit is transforming the public sector.* Reading, MA: Addison-Wesley.

Perry, J. L., & Kraemer, K. L. (1993). The implications of changing information technology. In F. J. Thompson (Ed.), *Revitalizing state and local public service: Strengthening performance, accountability, and citizen confidence* (pp. 225–245). San Francisco: Jossey-Bass.

Rahm, D. (1999). The role of information technology in building public administration theory. *Knowledge, Technology, and Policy, 12,* 74–83.

Reschenthaler, G. B., & Thompson, F. (1996). The information revolution and the new public management. *Public Administration Research Theory, 6,* 125–143.

Scavo, C., & Shi, Y. (1999). World Wide Web site design and use in public management. In G. D. Garson (Ed.), *Information technology and computer applications in public administration: Issues and trends* (pp. 246–266). Hershey, PA: Idea Group.

Sprecher, M. H. (2000). Racing to e-government: Using the Internet for citizen service delivery. *Government Finance Review, 16,* 21–22.

UN & ASPA (2001). *Benchmarking e-government: A global perspective—assessing the UN member states*. Retrieved February 5, 2003, fromhttp://www.unpan.org/egovernment2.asp.

West, D. M. (2000). *Assessing e-government: The Internet, democracy, and service delivery by state and federal government*. Taubman Center for Public Policy at Brown University. Retrieved February 5, 2003, from http://www.brown.edu/Departments/Taubman_Center/polreports/egovtreport00.html.

West, D. M. (2001). *State and federal e-government in the United States*. Providence, RI: Brown University. Retrieved February 5, 2003, from http://www.insidepolitics.org/egovt01us.html.

Electronic Commerce and Electronic Business

Charles Steinfield, *Michigan State University*

INTRODUCTION

The term electronic commerce came into widespread use after the first graphical Web browser—Mosaic—was developed in 1993 and freely distributed around the world. Drawn by the ease of use of the browser, millions of home consumers, businesses, and educators connected to the Internet, creating the conditions for Internet-based commerce. Businesses flocked to the Internet, attracted by the ease of setting up electronic storefronts and the potential access to a global market of Internet subscribers. Sales of goods and services to consumers, often referred to as business-to-consumer (B2C) e-commerce have grown steadily each year, despite the failure in 2000 and 2001 of so many new Internet businesses (known as dot-coms because of the domain name—".com"—in their Internet address). Even more dramatic has been the extent to which businesses have adopted the Internet for supporting exchanges with other firms, such as suppliers and business customers. This is generally called business-to-business (B2B) e-commerce. Because of this rapid growth, e-commerce has become an important subject of study in its own right, and many schools now offer courses and degree programs focusing on it.

CHAPTER OVERVIEW AND THEMES

This chapter examines the development of e-commerce and its influences on the way companies work with their suppliers, market their products to customers, and compete with old and new rivals. The historical development of e-commerce suggests that it emerged as a powerful force only after a truly open and standard data network—the Internet—was coupled with easy-to-use software—the graphical Web browser. Many innovative business models evolved in an attempt to take advantage of this new platform, coupled with a host of clever new marketing strategies. E-commerce opened up new avenues for consumers as well, creating new consumer-centric markets in which their bargaining power has been enhanced. New methods of payment as well as approaches to improving the security of online transactions have been introduced as well.

Despite these developments, a key theme of the chapter is that although much has changed as a result of e-commerce, many of the expected impacts have not occurred. A review of research on several anticipated impacts highlights the need to critically assess the early hype about the effect of e-commerce on competitive strategy, market structures, B2B relationships, and prices of goods and services.

The chapter concludes with a brief summary and a look toward what the future might hold now that the dot-com era of e-commerce has passed.

DEFINING E-COMMERCE AND E-BUSINESS

Today, the term electronic (or e-) commerce most commonly refers to the process of buying and selling goods and services over the Internet. It has even found its way into popular dictionaries, and the 2000 edition of the *American Heritage Dictionary* defines e-commerce as "[c]ommerce that is transacted electronically, as over the Internet." Narrower definitions state that complete transactions, including ordering and payment, occur entirely via the Internet, while broader definitions include information exchange in support of transactions, even if the actual payment occurs outside the Internet (often called “offline” payment). More recently, the term e-business has become popular. It refers more broadly to the conduct of business over computer networks. E-business includes activities that are not purely commercial transactions, such as when two firms use the Internet to collaborate on product development or research, or a firm provides customer service online. Some (e.g., Laudon & Traver, 2001) consider e-business to include only internal applications of a firm’s computer network, and not transactions with other firms. However, today, the terms e-commerce and e-business are often used interchangeably, and we follow this practice in this chapter.

A BRIEF HISTORY OF E-COMMERCE

Long before the World Wide Web, electronic networks were used to support transactions between businesses and their various external constituents (e.g., suppliers or customers). In the 1970s, forward-thinking manufacturers and wholesalers deployed proprietary data networks and simple terminal-based remote ordering systems to business customers. In three of the most famous examples from this era, pharmacists could replenish prescription drug stocks by filling out an electronic form on a terminal provided by McKesson, hospitals could use computer terminals to order a wide range of medical supplies from American Hospital Supply, and travel agents could look up flights on a computer terminal connected to American Airlines’ Sabre system. Soon the value of computer networks to link business buyers and sellers was well established, and efforts to create standard electronic documents to support trade were occurring across many industries. These latter standards were collectively called EDI (electronic document interchange). In some industries, such as automobile manufacturing and chemical production, EDI transactions proliferated because their use was mandated by large, dominant manufacturers. However, because EDI standards were complex and the proprietary data networks on which they relied were costly to implement, especially for small businesses, the extent of EDI diffusion was limited. Unlike the Internet, these proprietary networks took substantial time, investment, and effort before any new trading partner could join.

There were precursors to consumer-based e-commerce as well. Indeed, the Home Shopping Network on cable television, in which consumers can use a standard telephone to order goods displayed on TV, certainly resembles a Web-based electronic retailing model. Even more similar, however, were the many different electronic information services, collectively known as videotex, that developed largely in Europe and the United States in the 1980s. The most successful of these systems in Europe was the French system popularly known as Minitel, named after the small terminal that initially was freely distributed to telephone subscribers. Beginning in 1983, French telephone subscribers could use their Minitels to look up a wealth of information, engage in home banking, and order a wide range of goods and services, such as tickets, groceries, and other consumer products. As with the Internet, these services, offered over a public network to which anyone could connect, used a single standard and relied extensively on graphical content. Unlike with the Internet, a clear payment model was implemented, with France Telecom (at the time, the public administration responsible for the provision of all telecommunications networks and services in France) providing a “billing and collection” service for all companies that wished to sell via the Minitel. Consumers received the bill for their Minitel use on their regular phone bill, and France Telecom, in turn, paid the various service providers. Hence, the cost of participating in the Minitel marketplace was quite low, and by the early 1990s, there were over 25,000 services available, including both consumer-oriented and B2B services. Today, even with the Internet firmly entrenched in France and with many other telecommunications operators competing with France Telecom, the Minitel survives and generates revenue.

The success of the Minitel system did not go unnoticed, and many countries tried to create similar systems, including the United States. However, American electronic information services were still based on closed, rather than open, systems, using what has been called a “walled garden” approach. That is, services such as CompuServe and America Online (which have now merged) used their own proprietary networks, and their content and services were only available to their own subscribers. Unlike the Minitel network, it was not easy for any business to “join” the electronic market without significant investment. It was not until commercial traffic was permitted on the Internet, and an easy-to-use graphical browser appeared, that a truly open electronic marketplace began to grow and e-commerce as we know it today was born.

The early years of e-commerce had all the characteristics of a gold rush, as companies flocked to the Web to set up their electronic storefronts. Lured by the rapid growth of such first movers as Amazon.com, an enormous flow of investment capital was poured into Internet start-up firms selling everything imaginable on the Web. Laudon and Traver (2001), for example, report that more than $125 billion was invested in the initial public offerings (IPOs) of Internet firms between 1996 and 2000. The birth of a new digital economy was proclaimed, characterized by a seemingly unending growth in productivity rooted in the supposedly frictionless commerce afforded by the Internet. The new Web-based businesses theoretically offered such benefits as access to a larger potential market, lower inventory and building costs, more flexibility in sourcing inputs, improved transaction automation and data mining capabilities, an ability to bypass intermediaries that added costs but little value, lower costs for adjusting prices (known as a menu costs), increased ease of bundling complementary products, 7X24 access, and no limitation on depth of information provided to potential customers (Steinfield, Mahler, & Bauer, 1999). These advantages were expected to allow Internet firms to enter new markets at a lower cost and respond rapidly to shifts in market demand. Moreover, because the Internet also lowered consumers search costs (e.g., they did not have to drive from business to business to compare prices), prices would decline and Internet firms would prosper at the expense of traditional firms.

Despite the rosy projections for e-commerce, in late 2000 and early 2001, many of the dot-com companies began to fail. Few ever made any profits, as their business approach revolved around scaling up to increase market share. This approach was based on the belief that Internet businesses would enjoy positive network externalities—implying that the more people who used a site, the more value it would have for each user. The dot-coms that could attract the most users the fastest would, thus, have an insurmountable competitive edge. In search of the required scale, Internet firms kept prices too low and quickly burned through the cash raised from venture capitalists. At the same time, the U.S. economy began to slip into a recession, and venture capitalists began to lose faith in the viability of start-ups that failed to return any profits. Stock prices, which were far too high by normal investment standards, dropped precipitously, and the venture capital community lost interest in dot-com businesses. By the end of 2001, the dot-com era was over, and according to one estimate, only about 10% of the dot-coms created since 1995 still survived (Laudon & Traver, 2001).

B2B-oriented dot-coms experienced much the same fate. Large numbers of B2B electronic marketplaces were created to help match buyers and sellers in many industries, and by early 2000, more than 750 B2B e-markets were operating worldwide (U.S. Department of Commerce, 2000). Yet most of the new dot-coms in this arena failed as well.

Despite the problems of the dot-coms, statistics on e-commerce use do show steady gains, particularly as traditional companies have moved to enhance their offerings with Internet-based sales channels. According to U.S. Census figures, even as the economy was entering a recession and the dot-coms were failing, retail e-commerce sales in the United States were growing. Total U.S. retail e-commerce sales in 2000, the first full year that the government tracked this statistic, were approximately $29 billion. In 2001, this figure had increased to $36 billion, rising to more than $11 billion in the all important fourth quarter, where holiday shopping occurs (U.S. Census, 2002). Moreover, these figures are a low estimate, because they do not include sales from online travel, financial brokers, and online ticket sales. U.S. B2B online trade has also grown, rising to an estimated $466 billion in 2001 and projected to be as high as $5.4 trillion in 2006 (Laudon & Traver, 2001).

INTERNET BUSINESS MODELS

Many innovative methods of doing business were created in the early years of e-commerce, which collectively came to be called Internet business models. Exactly what constituted a business model was the subject of much debate, as was the issue of which models were successful. Rappa (2002) defines a business model as the "method of doing business by which a company can sustain itself—that is, generate revenue. The business model spells-out how a company makes money by specifying where it is positioned in the value chain." Timmer (1998) considers a business model to include three basic components: the architecture for the product, service, and information flows; the potential benefits for the various business actors; and the sources of revenues.

A common B2C business model, for example, is a virtual merchant that operates much as a traditional retailer, only without any physical retail space. The virtual merchant acquires products from manufacturers and resells them over the Web to end consumers. Amazon.com, which sells books, CDs, videos and many other consumer products online, is the best known example of a virtual merchant. Other popular B2C models include catalogue companies and the click-and-mortar or brick-and-click retailers, which have both traditional physical stores and online channels. All these businesses typically derive revenues from the margin between the prices that suppliers charge and the prices charged to consumers. However, each adds value in different ways to attract buyers. For example, a virtual merchant (also sometimes called a digital pureplay) may be able to offer goods at lower prices because of its lower inventory holding costs and lower selling costs. On the other hand, a brick-and-click firm may use its Web site to help generate more traffic in its physical stores by offering coupons or searches of in-store inventory (Steinfield, Bouwman, & Adelaar, 2002).

Early in the e-commerce boom years, many felt that revenue from advertising would sustain e-commerce. This made some sense for content firms, such as online newspapers, magazines, radio stations, and other media-oriented companies that were accustomed to reliance on sponsors to bring down the costs for their viewers, listeners, and readers. However, one failed Internet business model, pursued somewhat by Buy.com, involved the use of advertising revenue to subsidize product sales, so that consumers could buy products at heavily discounted prices.

Another problematic business model in the B2C arena involved manufacturers of products selling directly to end consumers and bypassing wholesale and traditional retail channels. Levis, for example, attempted to sell its blue jeans directly to online shoppers. Existing retailers, however, threatened retaliation. Given the relatively small percentage of sales generated by the Internet, Levis soon gave up on the direct-sale model and now uses its Web site mainly to promote sales of Levis at traditional retail outlets or to sell online through its retailers' Web sites.

Among the more popular B2B business models was the B2B electronic hub (Kaplan & Sawhney, 2000). B2B e-hubs were established in a variety of industries by third-party market makers. Many were vertical markets such as eSteel, which attempted to create an online marketplace for all steel industry raw materials and products. These are called vertical markets because they bring together firms operating at various stages inside a particular industry value chain. Others were horizontal, such as Ariba or Grainger, which offer online catalogues that provide buyer firms with access to suppliers of a range of materials needed to operate a business. They are called horizontal markets because they focus on products needed across many different industries and bring together buyers and sellers of a wide range of maintenance, operation, and repair goods (MRO). The basic business model is oriented around the provision of an electronic brokerage service, with the market operator deriving revenue from commissions on sales, and perhaps fees charged to sellers in order to have access to the market.

Some brokerage models operated to bring consumers together, as in electronic classified services and auction services, such as eBay. These consumer-to-consumer (C2C) businesses clearly added value by attracting large numbers of buyers and sellers, increasing the chances of finding viable matches. As with B2B hubs, their revenue model usually hinged on taking a commission on sales, or charging fees to sellers.

Rappa (2002) offers a useful summary of the many business models encountered in e-commerce. His main model categories include

Brokerage: Various forms of electronic intermediaries that bring together buyers and sellers. EBay, which provides C2C auction services, is probably the most famous example of an online broker model.

Advertising: Companies that provide content in the hope of attracting enough viewers that they can sell advertising. Virtually all online newspapers and magazines that do not charge subscription fees are examples of this model.

Infomediaries: Companies that offer something of interest to consumers (normally content) in return for customer information that they can sell to third-party marketers. The *New York Times* online service uses this model.

Merchant Models: Retailers that buy goods and resell them via the Internet. Amazon is the most famous online virtual merchant. Firms that have both traditional retail outlets and online sites are using another variation of a merchant model, often called a click-and-mortar or brick-and-click model. In the same product category as Amazon we find Barnes and Noble, one of the most famous brick-and-click merchants.

Manufacturer: Producers of products that sell directly to end consumers via the Internet. Apple and Dell, for example, both sell their computers through online stores.

Affiliate: Companies that derive revenue by referring customers to other e-commerce firms, who then provide a sales commission. Many online merchants, such as Amazon, work with thousands of affiliate partners, who provide purchase opportunities in return for a small commission. In addition, firms such as BeFree help merchants establish large affiliate networks to enhance sales.

Community: Sites where the content is provided by the users themselves and revenues mainly come from advertising. ExpertCentral.com, a site that relies on volunteers to answer questions on a wide range of topics posed by online visitors, is an often-mentioned example of this type of model.

Subscription: Sites that offer access to content and services for a flat fee over some period, like traditional magazine or newspaper subscriptions. The *Wall Street Journal* operates their online paper with this model.

Utility: Sites that sell content or services on a metered, or per-use, basis. This is an evolving business model that might be used in the new approaches to software distribution.

There are many variations on each of these basic forms, with many strategies for adding value within each type. Moreover, many Internet companies represent combinations of these models, rather than existing purely within one category or another. A basic problem is that most business models emphasize the sources of revenue and are not necessarily recipes for profitable businesses. In a sense, many dot-coms following these models forgot to include a profit model. Many spent so much to scale up operations or acquire customers that they lost money on every sale. This resulted in the ironic situation that the more revenue they earned each year, the larger the total loss!

MARKETING STRATEGIES FOUND IN E-COMMERCE

Electronic commerce offers many new opportunities for firms to interact with customers. It enhances firms' abilities to offer many customer services in a cost-effective manner. Moreover, e-commerce companies have experimented with a wide range of innovative marketing strategies made possible through software, data mining, and other IT capabilities. Several of the more popular e-commerce marketing techniques are briefly highlighted below.

Personalization

An e-commerce vendor can personalize its interaction with customers, creating what some call a one-to-one marketing relationship. Personalization generally requires that the online shopper provide some information about his or her interests or preferences to the vendor. This information is then used to tailor the Web site so that it reflects these interests. Cookies are often used rather than requiring a login to identify particular customers, and each customer then receives his or her own dynamically constructed page. Yahoo and other search engines and portals were among the first to develop these techniques. More elaborate techniques use purchase histories, clickstream information, and other data to personalize e-commerce services.

Customization

Many e-commerce vendors offer customers a "build-to-order" option that allows some degree of customization of the products to suit buyer tastes. This is common for computer vendors, but other types of e-commerce companies have also used e-commerce to offer this value-added service to customers. Lands' End, for example, offers hem-to-order pants. Even Levis, in their initial foray into e-commerce, attempted a "made-to-measure" blue jeans product.

Upselling and Cross-selling

Using simple automated rules as well as more sophisticated data-mining techniques, merchants can enhance their ability to both upsell and cross-sell to customers. Upselling involves convincing customers to trade up to a more expensive option or to add a complementary product to an order. Cross-selling involves suggesting other products that customers might like that are similar to the ones in an existing order.

Collaborative Filtering

Many e-commerce sites use a data mining technique known as collaborative filtering to provide better product recommendations to customers, enabling even more powerful cross-selling. Essentially, the purchases of other customers who have similar profiles are used as a source of recommendation. Amazon was one of the first to implement such a service. Whenever a customer selects a particular book, he or she is presented with a list of related books to consider. These recommended books were purchased by other shoppers who had also bought the initial book selected by the customer.

Rich Content

Unlike products on a shelf in a traditional store, products in an e-commerce store can be accompanied by in-depth content, including product reviews, detailed specifications, tutorials, testimonials, multimedia depictions of it in use, music, and so forth.

Affiliate Programs

Vendors can use the wider Web community to help sell products. This involves more than simple reciprocal linking and can include payments to Web affiliates who refer customers who purchase products. Amazon was a pioneer in this area as well, allowing virtually anyone with a Web site to become an Amazon affiliate. For example, an author might provide a link from his or her personal Web site to the book's purchase page on Amazon. The link is coded to identify the referrer, who then earns a slight commission on any purchase originating at the referring site.

Incentive Programs

In an effort to inspire greater customer loyalty, many e-commerce sites followed the lead of airlines and offered various types of incentive marketing programs. Frequent shoppers could build up points that entitled them to discounts and other benefits. Because all transactions are automated and it is so easy to collect information via the Web, these types of programs became much more accessible to all online merchants. As with airline frequent flyer programs, these techniques create switching costs, because customers who do not return know they will lose accumulated benefits.

Direct E-mail

E-mail, the most popular application of the Internet, is a valuable e-commerce marketing tool. Companies obtain email addresses of visitors and customers and regularly send targeted letters with new product information, announcements of sales, and other promotional content.

Viral Marketing

Increasingly, companies rely on customers to help spread the word about their products and services. Such a strategy is known as viral marketing, because the marketing material is carried by customers and spread throughout a target population. Often, dissemination is accomplished by giving users access to free services or games that require the participation of others. Paypal, for example, lets its customers send payments to others for free. Recipients must open an account with Paypal in order to transfer the funds into their regular bank account. This strategy has helped increase the number of Paypal account holders, which in turn entices more merchants to accept Paypal payments. Merchants, however, do pay a fee to receive funds.

Multichannel Marketing

Brick-and click-firms and catalogue companies capitalize on the synergies that arise from the integration of traditional and online channels. Traditional retailers, for example, can provide detailed product information online, immediate pickup in a store, and extensive support after the purchase online. Traditional catalogue companies can refer to their Web sites in every printed catalogue mailed to customers. In this way, they can gain the benefit of a regular reminder to look at their merchandise, with the cost savings that accrue when customers fill out online forms instead of calling a call center.

The various marketing strategies mentioned above are by no means an exhaustive list, but they do illustrate a number of possibilities to extend and enrich relationships with customers using network-based, computer-mediated transactions.

CONSUMERS IN E-COMMERCE

Up to now, the discussion has focused on mainly on e-commerce companies. However, e-commerce also has implications for the way in which consumers interact with each other and with businesses. One of the most successful, and profitable, e-commerce businesses is eBay, which began as an auction broker for C2C transactions, providing the equivalent of a nationwide online garage sale. However, unlike real garage sales, buyers could not physically inspect merchandise, and there was no process equivalent to handing over the item in return for cash. eBay's innovative solution to this problem was to develop a feedback mechanism whereby buyers and sellers rate each other after each transaction. Buyers rate the promptness with which sellers actually delivered the purchased items, the quality of the items, how well it was packaged, the accuracy of the online description, and anything else that might help future buyers decide whether to do business with this particular seller. Sellers also rate buyers, scoring them well if they pay promptly and deal honestly. Privacy is protected because people use pseudonyms. There is an incentive to behave properly, because buyers and sellers develop a reputation (hence, these are often called reputation systems). If someone develops a bad reputation, then he or she can only return to the system under a new pseudonym, and others may be less willing to deal with someone with no reputation.

eBay relies on an auction pricing format (although buyers now have the option to click on a "buy it now" button and forego the bidding process). Unfortunately, even with the elaborate reputation system, there remains some potential for fraud. Some unscrupulous sellers, for example, may fake numerous transactions with others to build a reputation. Others may have confederates make false bids on an item in order to artificially raise prices—this practice is known as shilling. Bidders can also cheat—for example by using a confederate who makes a high bid to discourage other bidders. The fake bid is then retracted near the bidding deadline, so that the item can be acquired at a lower price. Even with a certain degree of fraud, however, eBay remains highly successful, perhaps in part due to the entertainment value associated with bidding on items.

Consumers play an important role in another well-known e-commerce service, run by Priceline.com. Priceline also uses a form of electronic brokerage, in which they match consumers' bids for various goods and services with vendors willing to accept the bid. Priceline mainly deals with travel-related services, especially airline tickets, although they attempted to expand this model to other sectors. In this way, airlines and others with perishable products have another outlet for goods and services that would otherwise go unsold. This business approach is called a reverse auction broker model and is sometimes referred to as C2B e-commerce because the consumer initiates the transaction.

PAYMENT SYSTEMS TO SUPPORT E-COMMERCE

To support complete transactions online, the e-commerce infrastructure must be capable of handling the financial settlement. Since the dawn of e-commerce, many financial and technology companies have worked to establish secure methods of electronic payment. We can group most forms of electronic payments into three broad categories: (a) closed user group systems, (b) systems that use secure transmissions of traditional payment methods, and (c) token-based systems for digital money. Each of these is briefly described below.

Closed User Groups

A closed user group approach relies on payment occurring outside of the Internet. In general, this approach is used when online shoppers have an account and are billed on some periodic basis for their usage and/or purchases. For example, an AOL subscriber has made arrangements to pay AOL via a check each month, or via his or her credit card. Any online purchases can then be appended to this regular bill. Likewise, a customer may have a private account with a vendor that offers products online. The customer then logs in with his or her account ID and password, purchases items, and receives a bill at the end of the month for all activity on the account. These are both closed groups—only members are able to make purchases, and a billing relationship is already established, so that payment happens offline.

B2B exchanges often take place within a closed user group. For example, it is common for suppliers to have long-term contracts with their buyers for prenegotiated prices. Supply-chain management systems can permit online ordering of needed supplies, whereas payment occurs through the normal offline billing and account management process.

Methods of Secure Transmission of Existing Payments

Many e-commerce vendors sell products to new customers who do not have accounts, and hence a closed user group approach would not work. Rather, the most common method of payment in e-commerce is to set up a secure transmission link to accept an existing form of payment, such as a credit card or an electronic check. One elaborate standard, known as the secure electronic transaction (SET), was developed and supported by credit card firms, including Visa and Mastercard. However, the system so far has proved to be too costly and cumbersome, and most sites simply use some variation of the original secure sockets layer (SSL) approach to enable shoppers' browsers and e-commerce servers to set up secure links. Credit card information is encrypted to prevent interception and theft as it travels over the Internet. Merchants must verify the credit card, and consumers, of course, must pay their credit card bill through normal offline means each month.

B2B payments may also occur using secure transmissions, even when the companies are not in a closed user group, as described above. Secure transmission of electronic checks, such as is offered by eCheck, has been promoted as a safe means of transferring funds between businesses. Consumers may also use e-checks for their payment needs.

Token or Digital Money Systems

One problem with credit cards is that the fees charged by credit card companies make them less viable for inexpensive purchases. Content-oriented e-commerce companies are particularly interested in better methods of low-value payment. Some music sites, for example, might find it viable to charge just a few pennies to play a song, counting on generating millions of transactions so that the total translates into real money. In these cases, one solution is to have users set up accounts and then to provide a periodic bill that aggregates usage. However, such an approach sets up a barrier to spontaneous usage. Instead, researchers have attempted to develop digital money that could reside in an electronic wallet on a hard drive, or in a mobile device. The basic idea is to have the consumer download some funds into his or her computer. This electronic cash could then be spent online, with appropriate controls, so that once handed over to a vendor, it is properly debited from the shopper. Ideally, digital money could support micropayments (even fractions of a penny), enable anonymous transactions, and prevent anyone from using the same tokens to pay more than one recipient. A famous example of an electronic cash system is the now defunct eCash from a company known as Digicash. Most of the digital money systems have not met with much success in the market (McCullagh, 2001).

One of the most successful e-commerce payment systems actually combines some elements from the above types of systems. PayPal, a payment service provider used on such services as eBay, operates somewhat like a closed user group, a credit card system, and digital money. It operates somewhat like an online bank account, from which customers can make purchases from vendors who use it. They can put funds into their accounts using a credit card or through funds transfer from their regular bank account. But PayPal is more open than other closed systems and relies on more than simple secure transmissions of card numbers. As noted earlier, account holders can send money to anyone with an e-mail address. This is an excellent example of what is called a viral marketing strategy, because anyone receiving payment by PayPal has to open up an account to actually obtain the money.

EMERGING TECHNOLOGIES

New technologies are constantly being developed for e-commerce. Some focus on enhancing the richness and vividness of the e-commerce experience, in an attempt to make it as compelling as shopping in real life. Work on virtual and 3D environments is ongoing, in anticipation of the day when consumers will have computers with fast processors and 3D video cards coupled with broadband Internet access. Early approaches have been used by clothing catalogue companies, such as Lands' End, which allow shoppers to create 3D models of themselves that can try on clothes. Each shopper can modify the shape and appearance of his or her model and then get some idea of how particular articles of clothing might look without personally trying them on. Virtual reality approaches go one step further, placing the online shopper in a synthetic environment. This is ideal for tourism, real estate walkthroughs, museum visits, and other types of products or services requiring a more experiential selling approach.

Another emerging set of technologies involves the design of software agents that act on behalf of an online shopper or vendor. One early type of agent was known as a shopping bot (short for robot), which could search out items and provide price comparisons. Agents can be programmed to perform such tasks as track and place bids at auctions, find and negotiate prices on desired items, and buy and sell stocks. Research today explores how agents behave, especially when interacting with other agents in what are called multiagent systems.

Perhaps one of the most exciting new e-commerce developments to watch is the extension of online shopping into the mobile arena. Mobile handset manufacturers such as Nokia are building cellular phones with Web browsing capability. Next-generation cellular networks support packet switching and Internet protocol traffic. New protocols, such as the wireless application protocol (WAP) and the wireless markup language (WML), enable the quick translation of Web pages into a format readable on small cellular phone and PDA screens. Once this user and network infrastructure is in place, people will no longer be tied to a desktop computer in order to surf the Internet and shop online. They are free to engage in e-commerce anytime and anyplace. Mobile, or m-, commerce is generating much interest, primarily because the global number of cellular subscribers is now more than 1 billion, which far exceeds the number of people who have computers connected to the Internet.

Many m-commerce services will be based upon the particular location of shoppers, allowing someone in a car, for example, to find the nearest vendor of a particular product. These types of services are collectively known as location-based services. In the United States, the FCC has mandated that all cellular operators be capable of determining the location of cell phone users in order to support emergency 911 services. However, there are numerous e-commerce uses of location data, and we are likely to see a wide range of services appear. Users may use their mobile device to search for available vendors in an area, receive directions from their current location to a vendor, make payments at the point-of-sale device, and receive advertisements based upon their interests when they near particular locations. Location-based services thus blend e-commerce and physical commerce in new ways, blurring the traditional online/offline distinctions of the past.

BARRIERS TO E-COMMERCE DEVELOPMENT

E-commerce has grown rapidly, but still it accounts for only a small fraction of purchases due to a number of critical barriers. Of course, there are always economic and technological barriers that can explain low e-commerce usage in many parts of the world. Obviously, e-commerce, especially when directed at consumers, has little relevance in places where there is limited computer penetration and most households do not have phone lines. Yet even in countries like the United States—where 94% of households have telephone service, two thirds of households have computers, and more than half of the population uses the Internet—only about 1% of retail trade occurs online. Among the more frequent explanations for low take-up of e-commerce are lack of trust in online vendors, security concerns, and incompatible consumer needs and desires.

Trust was largely an issue when thousands of new and unfamiliar dot-com companies flooded the Internet. Consumers were reluctant to shop at unfamiliar online stores for fear that their orders would not be fulfilled, or that they would not be able to return defective merchandise. The same problem prevents businesses from buying from unfamiliar suppliers, who have not proved that they are both trustworthy and competent to handle the business. To combat this problem, e-commerce companies often relied on trusted third parties—firms like Verisign or Trust-E—that would vouch for the legitimacy of an online vendor. Dot-com companies display the icon from these trusted third parties to help generate trust. Today, this issue is abating as e-commerce brands become better known, and as established firms move into e-commerce. However, consumers have new concerns about how e-commerce companies will use the information they collect. This is a privacy issue on its face, but it also reflects a lack of trust that online vendors will behave responsibly. Many firms now proclaim their privacy policy to help combat this issue.

A related issue is the fear that many online users have over the security of online transactions. Theft of credit card numbers and personal information are chief concerns. Secure transactions help to some extent, but these do little good when hackers find ways to break into corporate networks and steal entire databases of credit card numbers stored on servers.

Finally, for many purchases, e-commerce may not be the best approach given consumers' needs, desires, or home situations. For example, when someone needs a product immediately, they are much more likely to pick it up at a local store than wait one or more days for delivery. Some click-and-mortar firms, such as Best Buy, recognize this need and now offer in-store inventory search and pickup options for online shoppers. Other consumers see shopping as a social and entertainment activity, and a visit to the mall with friends is hardly replaceable by e-commerce. Finally, some services simply require too much of consumers to be viable. For example, in order to buy groceries, including perishable items, online, someone needs to be home at the time of delivery. This scheduling may be difficult, and to get around it, some companies attempted to use refrigerators in garages to which deliverymen had access. Yet this also requires consumers to open up their private space to strangers, and the costs of installing appliances at customer premises proved to be impractical. Moreover, not everyone had a garage, which limits the market. These are just a sampling of shopping situations where e-commerce might not be the best option.

THE ECONOMIC AND COMPETITIVE IMPACTS OF E-COMMERCE

E-commerce represents both a threat and an opportunity for most companies. It can be expensive to implement, and firms must carefully evaluate the competitive benefits they might achieve. In this section, we examine some of the competitive and economic issues raised by the advent of e-commerce.

Competitive Advantage and E-business

Information systems researchers have long recognized the fact that information technologies are important competitive weapons (Porter & Millar, 1985). Innovative IT applications and networks help firms to lower costs and allow them to offer new value-added services that serve to differentiate a company from its competitors, which are the two basic competitive strategies discussed by Porter (1985). Information technologies and networks can be deployed throughout every firm's value chain, which is the set of activities through which any company converts inputs, such as raw materials and labor, into products that can be sold to customers. E-commerce can permeate the value chain, reducing the costs of acquiring supplies, enabling firms to reach into new markets via the Internet, and supporting such value-added services as product customization and personalization of the information made available to customers.

By enabling access to distant markets without requiring expensive brick-and-mortar investments, e-commerce lowers important barriers to the entering of new markets. Moreover, because it is so easy to bundle new products over the Web, e-commerce can introduce new competitors into existing industries. Hence, one effect of e-commerce is that incumbent firms in any industry have faced new competitors that were formerly kept out by virtue of geography or other high-cost factors. New technologies may also substitute for existing ones, such as when consumers use new music formats in place of purchasing CDs at stores or online.

Yet e-commerce has another effect due to its ability to help companies pursue a differentiation strategy. Companies that gather information from customers and use it to help offer highly personalized services, or that provide incentives to keep customers returning to their Web site, create what has been called "sticky" e-commerce (Shapiro & Varian, 1999). This is when e-commerce raises customers' switching costs (the costs to move to a different supplier of a good or service), raising a new barrier that can help prevent other firms from successfully attacking a particular market. Network externalities—which create positive feedback loops, so that the larger the service, the more attractive it is—also work to make it difficult for new firms to attack established players. For example, eBay is so attractive because its large user base increases the likelihood that people will always find the specific product they are seeking. This, in turn, draws more sellers, which attracts more buyers, and so on. Smaller services have trouble competing with this dynamic. Such differentiation effects demonstrate that e-commerce does not make exchanges frictionless, and this result would not, in any case, be in the interest of merchants.

Alternative perspectives point out other reasons why new entrants may find it difficult to dislodge incumbents in any industry. Afuah and Tucci (2001) point out that incumbent firms normally possess crucial resources that new entrants may not have. These are often called complementary assets, and they include such factors as established relationships with suppliers, experience, access to retailers, powerful information systems, and strong distribution networks. Indeed, the rise of click-and-brick e-commerce shows forcefully how incumbent retailers can better capitalize on e-commerce than the new dot-coms that do not possess the required complementary assets.

There are many other potential impacts of e-commerce on the competitive dynamics of industries. In addition to the rivalry among competitors in an industry, and the potential threat from new entrants and substitute products noted above, Porter (1980) directs our attention to threats arising from elsewhere in the value chain, including suppliers and buyers. A good example of e-commerce enhancing the bargaining power of suppliers occurs when manufacturers threaten to bypass retailers to sell directly to customers. To the extent that this threat is credible, it may reduce the ability of retailers to capture more of the value when they resell products and services to end consumers. The threat of disintermediation is discussed in more detail below. Good examples of the ways that e-commerce can enhance buyer power are noted in the section above, on consumers in e-commerce. Reverse auctions such as Priceline illustrate how a market can become more buyer-centric. Other enhancements to buyer power through e-commerce occur when individual consumers band together to obtain the types of volume discounts normally only offered to larger business buyers. Third parties that organize this cobuying capability are known as buyer aggregators.

E-commerce and Transaction Costs

The primary way that e-commerce helps companies reduce costs and enable their entry into new markets is by reducing what economists call transaction costs (Williamson, 1975). Every commercial exchange, including C2B and B2B exchanges, is a transaction that can be broken into a number of discernible stages, each with its own set of costs to the participants. At the simplest level, we can consider any purchase to require an information or prepurchase phase, a purchase phase, and an after-sales, or postpurchase, phase. During the information phase, for example, buyers face search costs, while sellers have costs in supplying purchase information to buyers. During the purchase face, other costs are associated with negotiation, monitoring contracts, and actual financial settlement. Postpurchase costs include repairs, returns, and other types of customer services. When transaction costs get too high, markets become inefficient, and prices can be higher than they should be. For example, imagine a small town with one seller of some product. Without the Internet, customers in this town face high search costs (e.g., driving to another city to look in stores there) with uncertain payoff. So they pay the higher prices charged by the local monopoly provider. With the Internet, customer search costs for finding lower-priced suppliers are dramatically reduced, forcing the local monopoly seller to lower its prices. Because of this effect, many economists believe that the Internet will help reduce prices across all industries (Bakos, 1997). Indeed, it became popular to talk about the "death of distance" as a barrier to commerce due to the reduced transaction costs afforded by e-commerce. However, such an analysis overlooks many important influences on the price of any product, including the fact that online products differ in many ways from their counterparts in a store. Online products are not immediately available and cannot be touched, smelled, or otherwise physically examined, and perceived risks (e.g., ease of returns) may be higher. It further ignores the importance of established relations between buyers and sellers, a notion discussed next. Transaction cost frameworks must be extended to deal with these issues.

Market Structure and Buyer–Seller Relations

The reduction in transaction costs afforded by e-commerce is proposed to have another big economic effect, primarily at the B2B level. If transaction costs can be reduced, then market mechanisms may be used where they were not feasible before, based on an early theoretical argument proposed by Malone and colleagues (Malone, Yates, & Benjamin, 1987). For example, suppose a company needed a very specific type of input, only available from one particular supplier in its area. In the past, this company was at a disadvantage. Due to information asymmetries (the supplier knew what it cost to make the goods, but the buying company did not) and lack of alternative providers, the supplier could charge higher prices. However, the reduced search costs and greater information available on the Internet reduce these asymmetries and increase choice. Hence, firms can avoid being locked into long-term relationships and can use more spot-market buying behavior to source the inputs they need. The theoretical prediction was that such uses of e-commerce would give rise to large electronic marketplaces where businesses could buy inputs based purely on the best available offers.

Although e-commerce did result in the formation of large B2B electronic marketplaces that offered both manufacturing (vertical) inputs and commodities (horizontal) supplies, the above prediction has not been entirely supported. First, most of the third-party-developed B2B markets have failed (Laudon & Traver, 2001). Substantial empirical evidence suggests that firms are more likely to use e-commerce to increase the efficiency of transactions with trusted suppliers than to enable spot transactions with new suppliers (Steinfield, Kraut, & Plummer, 1995). Hence, they were less likely to allow new third-party firms to position themselves as profit-making intermediaries in these well-established relationships. Instead, the bulk of B2B e-commerce occurs over private industrial networks typically organized by larger buyers or sellers. Often a large company will function as the leader in a value network that encompasses firms up and down the value chain. These value networks compete with other value networks, as an extension to simple firm-to-firm competition. The B2B e-commerce links permit more than just simple buy and sell transactions—they enable collaborative e-commerce among networks of firms, allowing such activities as codesign of components and joint marketing efforts.

Intermediation and E-commerce

As noted above, e-commerce can enable producers of goods and services to bypass intermediary firms, such as wholesalers and retailers, and sell directly to end customers. This effect has come to be called disintermediation and is often anticipated because of the belief that the Internet reduces producers' transaction costs for accessing end customers. Many producers were swayed by this logic and began to sell products directly from their Web sites. For some companies, such as Dell Computer and Cisco, the direct-sale business model made a great deal of sense, and the Internet provided a cost-effective way to manage it. Most of their customers were businesses already connected to the Internet and accustomed to doing business in this way. For others, especially those selling to consumers, this approach made less sense. Sarkar and colleagues showed early on that, in fact, the Internet provided as much or more opportunity for new forms of intermediation, and that it also helped to strengthen the role of existing intermediary firms (Sarkar, Butler, & Steinfield, 1995). They argued that intermediaries' transaction costs were also reduced, and the gains they get from e-commerce may negate any advantages producers might gain. They also note that intermediaries provide many functions to both buyers and sellers, such as needs assessment and evaluation, market information, and risk reduction, which are critical for many purchases. Many subsequent analyses demonstrated the fallacy of the disintermediation hypothesis (e.g., Schmitz, 2000). Marketing theorists caution that although particular intermediaries might be eliminated, their functions cannot be, which suggests that a role remains for intermediaries even in the age of e-commerce (Stern, Ansary, & Coughlin, 1996).

Pricing and E-commerce

Because of the reduction in transaction costs, and because of access to information about prices of alternative vendors, a primary effect of e-commerce was supposed to be lower and more homogenous prices relative to traditional channels. Up to now, the evidence for lower prices is rather mixed, especially when the added costs for shipping are taken into account. Moreover, price dispersion remains high, contrary to the expectations of economists (Smith, Bailey, & Brynjolfsson, 2000). It appears that because it is so easy to change prices (referred to as menu costs), Internet companies make more changes in smaller increments in response to the market.

Another reason there is variability in prices is that e-commerce makes it easier to engage in price discrimination, whereby vendors sell the same product to different customers at different prices. The goal, of course, is to learn what each customer's willingness to pay is, and then to charge him or her that price. One way this can happen is through auction pricing, which allows customers to submit their willingness to pay in the form of bids. Another popular price discrimination approach is to charge different prices for different versions of the same product or service (e.g., a student version of software versus a professional version) (Shapiro & Varian, 1999). Sometimes e-commerce companies have tried to estimate willingness to pay based upon some aspect of online behavior. Amazon.com, for example, once angered its customers when it was revealed that the online vendor had attempted to charge higher prices to returning customers, based on the assumption that these loyal customers would be less likely to switch to other online sellers. They soon stopped this practice. Another method of price discrimination is to make assumptions about willingness to pay based upon the referring site. Thus, if a shopper arrives at an online retail site by way of a comparison shopping agent or a discount shopping site, he or she may be presented with a lower price than a shopper who had clicked on an ad in an upscale investment journal.

Collectively, the research in this area shows that e-commerce effects are not so straightforward as was once believed. Many popular notions about e-commerce effects, such as the rise of disintermediation, the death of distance, and the emergence of frictionless commerce, are now viewed as myths.

CONCLUSION

In this chapter, we have provided a broad overview of e-commerce, introduced a brief definition, and showed how the creation of a ubiquitous, easy to use, and open data network helped e-commerce emerge as a powerful force from the many precursor systems and technologies. We discussed the development of innovative business models and noted the problems that dot-coms faced when they attempted to delay profits in exchange for market share. Many powerful marketing strategies were introduced, as was the rise of consumer-centric markets. Ancillary technologies needed to enable e-commerce, such as payment systems, and emerging technologies, such as mobile commerce that will extend e-commerce, were also discussed.

A critical theme of the chapter is that e-commerce impacts are not straightforward, and claims of revolutionary changes must be carefully scrutinized. Important barriers to its development remain. Research on the competitive and economic impacts yields many findings that run counter to early expectations.

Despite these barriers and unanticipated impacts, our analysis still suggests that e-commerce will become increasingly pervasive, even without the headlines that tracked every new dot-com in the early years. There are a number of reasons for this expectation. First, as noted earlier, e-commerce usage is continuing to grow, despite the economic slowdown plaguing much of the world in 2001 and 2002. Moreover, even though IT spending is down and telecommunications firms are currently struggling, the high pace of investment throughout the 1990s means that the infrastructure is now in place to support e-commerce. The Internet backbone networks are fully capable of supporting more traffic and can accommodate the slow but steady growth in broadband access. This will, we hope, improve the responsiveness of e-commerce sites, encouraging more use. New technologies, such as the extension of e-commerce to mobile devices, will likely increase the pervasiveness of e-commerce. The dot-com bust may cause some to be pessimistic, but it can also be viewed as a normal evolutionary process at the start of a new industry. Poor business models have failed, but viable ones have survived and may prosper once the economy improves. Moreover, the entry of established companies into e-commerce is lending more legitimacy to the sector and should help overcome trust problems. These firms also have the resources and alternative sources of income to enable their e-commerce channels to develop appropriately, without reckless moves aimed at capturing market share at the expense of profits. Finally, continued B2B e-commerce development is enhancing efficiency in the supply chain, which can also carry over to B2C e-commerce. In summary, e-commerce will continue to spread throughout the economies of the world, but our analysis suggests that the economic effects will be complex and, at times, counterintuitive.

GLOSSARY

B2B electronic hub Electronic marketplaces where business buyers can acquire goods from suppliers. Vertical hubs focus on the provision of manufacturing inputs bringing together firms operating at various stages inside a particular industry value chain. Horizontal hubs focus across industries, bringing buyers and sellers of a wide range of maintenance, operation, and repair goods (MRO) to the market.

Buyer aggregation The process of organizing small buyers into a cooperative buying group in order to obtain volume discounts.

Collaborative e-commerce A form of B2B e-commerce among cooperating networks of firms that involves more than just purchases; also, such activities as the codesign of components and joint marketing efforts.

Complementary assets Resources possessed by a firm that enable it to take better advantage of innovations like e-commerce. These resources include established relationships, product know-how, existing business processes and systems, and distribution systems.

Disintermediation An effect of e-commerce whereby producing firms bypass traditional intermediaries and sell directly to end consumers.

Dot-com A name given to Internet companies because their network name ended in ".com."

EDI Electronic document interchange, a standard for transmitting business documents over computer networks.

Intermediary An entity that facilitates trade between buyers and sellers. Wholesalers and retailers are considered to be intermediaries that help manufacturers sell their products to the end consumers. Real estate agents are intermediaries who bring together home buyers and sellers.

Internet business model A generic term referring to the methods of doing business and generating revenue on the Internet.

Network externalities A term describing the benefits generated by having more people who used a service, such that the more people who use it, the more value it has for all users.

Price discrimination A pricing strategy whereby vendors sell the same product to different customers at different prices in an attempt to maximize overall revenue by matching prices to customers' willingness to pay.

Switching costs The costs a buyer faces when changing to a different supplier of a good or service.

Transaction cost Various indirect costs faced by buyers and sellers as they complete a commercial exchange, including search and information-gathering costs, monitoring and settlement costs, and after-sales and service costs.

CROSS REFERENCES

See *Business-to-Business (B2B) Electronic Commerce; Business-to-Business (B2B) Internet Business Models; Business-to-Consumer (B2C) Internet Business Models; Click-and-Brick Electronic Commerce; Consumer-Oriented Electronic Commerce; Electronic Data Interchange (EDI); Electronic Payment; E-marketplaces; Internet Literacy.*

REFERENCES

Afuah, A., & Tucci, C. (2001). *Internet business models and strategies: Text and cases*. New York: McGraw-Hill Irwin.

Bakos, J. Y. (1997). Reducing buyer search costs: Implications for electronic marketplaces. *Management Science*, *43*(12), 1676–1692.

Choi, S. Y., Stahl, D., & Whinston, A. (1997). *The economics of electronic commerce*. Indianapolis, IN: Macmillan Technical Publishing.

Kaplan, S. N., & Sawhney, M. (2000). E-hubs: The new B2B marketplaces. *Harvard Business Review, 78*(3), 97–103.

Laudon, K., & Traver, C. (2001). *E-commerce: Business, technology, society*. Boston: Addison-Wesley.

Malone, T., Yates, J., & Benjamin, R. (1987). Electronic markets and electronic hierarchies: Effects of information technology on market structure and corporate strategies. *Communications of the ACM, 30*(6), 484–497.

McCullagh, D. (2001, June 14). Digging those digicash blues. *Wired*. Retrieved October 6, 2002, from http://www.wired.com/news/exec/0,1370,44507,00.html

Porter, M. E. (1980). *Competitive strategy: Techniques for analyzing industries and competitors*. New York: Free Press.

Porter, M. E. (1985). *Competitive advantage*. New York: Free Press.

Porter, M. E., & Millar, V. E. (1985). How information gives you competitive advantage. *Harvard Business Review*, *63*(4), July–August, 149–160.

Rappa, M. (2002). *Business models on the web*. Retrieved October 6, 2002, from http://digitalenterprise.org/models/models.html

Sarkar, M., Butler, B., & Steinfield, C. (1995). Intermediaries and cybermediaries: A continuing role for mediating players in the electronic marketplace. *Journal of Computer Mediated Communication, 1*(3). Retrieved October 6, 2002, from http://www.ascusc.org/jcmc/vol1/issue3/vol1no3.html

Schmitz, S. (2000). The effects of electronic commerce on the structure of intermediation. *Journal of Computer Mediated Communication, 5*(3). Retrieved October 6, 2002, from http://www.ascusc.org/jcmc/vol5/issue3/

Shapiro, M., & Varian, H. (1999). *Information rules: A strategic guide to the network economy*. Boston: Harvard Business School Press.

Smith, M., Bailey, J., & Brynjolfsson, E. (2000). Understanding digital markets: Review and assessment. In E. Brynjolfsson & B. Kahin (Eds.), *Understanding the digital economy* (pp. 99–136). Cambridge, MA: MIT Press.

Steinfield, C., Bouwman, H., & Adelaar, T. (2002). The dynamics of click and mortar e-commerce: Opportunities and management strategies. *International Journal of Electronic Commerce, 7*(1), 93–119.

Steinfield, C., Kraut, R., & Plummer, A. (1995). The effect of networks on buyer-seller relations. *Journal of Computer Mediated Communication, 1*(3). Retrieved October 6, 2002, from http://www.ascusc.org/jcmc/vol1/issue3/vol1no3.html

Steinfield, C., Mahler, A., & Bauer, J. (1999). Electronic commerce and the local merchant: Opportunities for synergy between physical and web presence. *Electronic Markets*, *9*(1), 51–57.

Stern, L. W., El-Ansary, A. I., & Coughlan, A. T. (1996). *Marketing channels*. Upper Saddle River, NJ: Prentice Hall-International.

Timmer, P. (1998). Business models for electronic markets. *Electronic Markets, 8*(2), 3–8. Retrieved October 6, 2002, from http://www.electronicmarkets.org/netacademy/publications.nsf/all_pk/949

U.S. Census Bureau (2002). *Service sector statistics*. Retrieved October 6, 2002, from http://www.census.gov/mrts/www/current.html

U.S. Department of Commerce (2000). *Digital economy 2000*. Retrieved October 6, 2002, from http://www.esa.doc.gov/508/esa/DigitalEconomy.htm

Williamson, O. (1975). *Markets and hierarchies: Analysis and antitrust implications*. New York: Free Press.

Electronic Data Interchange (EDI)

Matthew K. McGowan, *Bradley University*

INTRODUCTION

Electronic data interchange (EDI) is the computer-to-computer exchange of business transactions in standardized formats. EDI represents a type of interorganizational information system (IOIS) because it involves an electronic link from one organization to another organization. It is also one of the earliest forms of electronic commerce. Businesses use EDI to conduct business-to-business (B2B) electronic commerce. B2B electronic commerce is the largest portion of electronic commerce. According to a recent study by the United States Census Bureau, manufacturers were the largest business users of electronic commerce, conducting 18.4% of their orders as electronic transactions in 2000. EDI allows businesses to send and receive business transactions electronically, resulting in faster processing and greater accuracy. EDI is an enabling technology for full business-to-business integration of systems. Bell Helicopter (2002) uses EDI to support many aspects of its business with suppliers, including request for quotes, shipping schedules, order status, invoices, and payments.

EDI requires (1) translation of transactions into standard EDI formats, (2) transmission of transactions to the trading partner, and (3) translation from the standard formats into the formats used by the information systems processing them. In traditional EDI, companies frequently rely on communication service providers for private network communications links with trading partners. Internet-based EDI is rapidly evolving, permitting companies to bypass private networks in favor of the Internet. Internet-based EDI has the potential to reduce communication costs but carries higher security risks than traditional EDI. Large companies may be able to provide Web-based alternatives to traditional EDI for their small trading partners that require little technical expertise or expense.

Potential disadvantages of EDI include security, transaction errors, authentication, nonrepudiation, reliance on third parties, and the costs of implementing EDI. The technical issues of implementing EDI may seem minimal to a large company with an EDI specialist, but can seem overwhelming to small companies with little technical expertise (McGowan & Madey, 1998).

Many organizations credit EDI with attainment of strategic and tactical advantages. EDI may also present organizational and administrative challenges. Companies must develop agreements with their trading partners and consider whether this information sharing will affect their bargaining power with their trading partner(s).

The following section contains a description of EDI and provides some historical perspective on its development.

WHY EDI?

Organizations have traditionally relied on paper in conducting business. Many companies developed standard forms to organize the information needed to conduct a transaction and to be sure that the information was complete. The postal service was a primary means of exchanging business documents. Innovations such as express delivery services (e.g., FedEx) and facsimile transmission allowed organizations to conduct business transactions more rapidly. Paper, however, remained the primary communications medium. When an organization receives a business document by fax or paper, someone must enter the information from the form into the appropriate information system. This step takes time and introduces opportunities for human errors.

To send a business form to a trading partner, the organization must first print the document from the information systems and then mail or fax it to the trading partner. Each organization is likely to use a different format for its forms. Because the forms do not have a standard format, it may be difficult to obtain the appropriate information from the form.

EDI, on the other hand, uses standard formats, which makes it possible to parse the information into a format for the appropriate information system to process. The business document exists electronically, so software can convert the document into a form suitable for the business application and no human intervention is required.

A company could send business documents via traditional electronic mail. That would still require human intervention and would still lack a standard format. There is the potential for using electronic mail for EDI so that no human intervention is required. A subsequent section of this article dealing with Internet-based EDI addresses that option.

HOW EDI WORKS

What follows is a discussion of how EDI operates, the alternative communications approaches for EDI, the costs of EDI, and the role of Value Added Network (VAN) providers.

Overview

Electronic data interchange uses a standard format for the exchange of business documents between two trading partners. Figure 1 illustrates an example of the EDI process. The example shows a supplier sending an advance shipping notice to a buyer. Specifically, the supplier's shipping system is sending advance shipping notices to a buyer's receiving system. In this example, the data already reside in a computer information system. The first step in the EDI process is for an application program to extract the data from the shipping system. Next, a program translates the data into the EDI standard format for an advanced shipping notice. In the third step, EDI translation software converts the extracted data into the standard format for EDI. The supplier's EDI document management software transmits the document(s) to the buyer via a communications network.

The communications network can be a private network such as a value added network (VAN) or it can be the public Internet. The buyer receives the advance shipping notice via its EDI document management software and translates the documents into a file for the appropriate application system (in this case, the receiving system), and then an application program in the receiving system loads the documents into the buyer's receiving system.

Organizations that do not have the data for EDI transactions in existing computer systems can still use EDI to exchange business transactions. Several EDI software vendors sell software that permits a company to enter data directly into the EDI software. The software then formats the EDI transactions and transmits them over the communications network. A company could even choose to print the data from an information system and rekey it into the EDI software, rather than create an extract program. Similarly, companies can receive transactions into EDI software and then print the business documents from the EDI software. People frequently refer to this form of EDI as paper-based EDI because of the paper involved with the process. However, companies do not obtain the benefits of a fully integrated system when using this technique (McGowan & Madey, 1998).

Value Added Network Services

VAN service providers can assist with many aspects of the EDI process. VAN services can include electronic

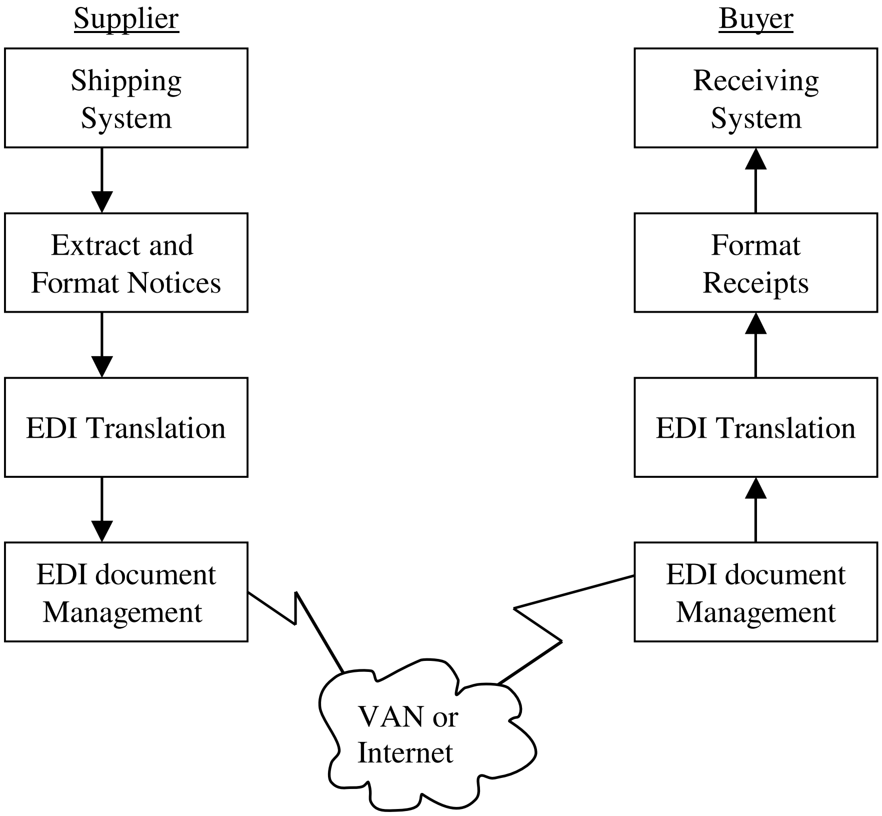


Figure 1: Steps in the EDI process.

mailboxes, connectivity, translation services, security, and an audit trail (Copeland & Hwang, 1997). The electronic mailbox feature is the most common reason for using a VAN. The VAN stores EDI documents into the company's electronic mailbox, allowing a company to retrieve the documents as desired. The VAN can also provide a service that forwards documents to the company or notifies the company if the company desires immediate notification or uses event-driven processing in its business (Copeland & Hwang, 1997).

The VAN provides connectivity to other trading partners. The VAN can store a company's documents in the electronic mailboxes of its partners, or it can forward a document to a partner's VAN (Copeland & Hwang, 1997). Thus, a company does not need to maintain a mailbox with each one of its partners' VAN providers.

VAN services can include translation of data into EDI formats. A company could provide an extract of its business data to the VAN and the VAN could translate the data into EDI documents.

A VAN provider can help with security. Security services include access control, authentication, and encryption (Copeland & Hwang, 1997). Mailbox access generally requires a user ID and password. The VAN can check that an authorized trading partner sent documents and that the documents are in a valid format. Encryption service can ensure confidentiality. VAN providers can provide private line connections or switched access to avoid potential security flaws of the Internet.

The VAN may also provide an audit trail for documents processed via EDI. The VAN can provide automatic acknowledgments to trading partners that require them. The VAN can also provide tracking information for documents (Copeland & Hwang, 1997).

Transmission Alternatives

There are three common alternatives for transmitting EDI documents: private VAN network(s) the public Internet, and direct connection to the trading partner. Traditional EDI uses a VAN network service to send the EDI documents to the trading partner. The VAN stores the EDI documents in an electronic mailbox that it manages for the trading partner. If the trading partners use different VAN providers, then the VAN provider of the sending trading partner forwards the EDI documents to the VAN provider of the receiving trading partner.

The second alternative for transmitting EDI documents is Internet-based EDI. Several techniques are available for Internet-based EDI and it is the fastest-growing form of EDI. All Internet-based EDI techniques rely on the public Internet as the communications network. A more complete discussion of this popular form of EDI will be presented later in this article.

The third alternative for transmitting EDI documents is for one trading partner to maintain a proprietary electronic mailbox system. With the proprietary system, the company maintaining the system allows its trading partners to send documents to and receive documents from the electronic mailboxes. In this case, the EDI traffic is limited to transactions involving the company hosting the proprietary system. A large company might use this approach to reduce the charges from its VAN provider.

Companies may use several of these approaches for their EDI transactions. For example, a company might use traditional EDI for some trading partners, Internet-based EDI for other trading partners, and proprietary systems for a third group of trading partners.

EDI STANDARDS

The EDI process requires that business documents be in a standard format. The first instances of EDI relied upon proprietary formats to exchange business documents. Wal-Mart and K-Mart used proprietary formats when they first implemented EDI (Choudhury, 1997). The problem with proprietary standards is that it is difficult to exchange transactions with many trading partners because of all the format conversions (Chan, 2002). Having standard formats for business documents means that all trading partners can understand the document structure and interpret it correctly.

There are two major sets of standards for EDI: the American National Standards Institute Accredited Standards Committee X12 (ANSI ASC X12, or more usually, X12) standard, and the United Nations Electronic Data Interchange For Administration, Commerce, and Transport (EDIFACT) standard.

The EDIFACT Standard

The United Nations developed EDIFACT as an international standard to help facilitate EDI for international commerce. The International Standards Organization (ISO) and the United Nations Economic Commission for Europe share responsibility for developing and maintaining the EDIFACT standard (Chan, 2002). The EDIFACT standard is essentially a combination of the X12 standard from the United States and the Trade Data Interchange (TDI) standard developed in the United Kingdom (Chan, 2002). The EDIFACT standard includes components similar to the X12 standard.

The X12 Standard

ANSI specified the development of the X12 standard to help facilitate EDI in the United States, and it is the most commonly used format in the United States. The X12 standard defines the data structure, content, syntax, and sequencing for business documents. The X12 standard specifies a type of business document as a transaction set. For example, there are transaction sets for business documents such as Purchase Order, Invoice, and Advanced Shipping Notice. There are currently more than 275 different transaction sets in the X12 standard. The standards committee assigns a name and number to each transaction set. Table 1 shows a partial list of the transaction sets. For example, an invoice document has the code of 810.

The X12 standard specifies data segments and data dictionary for each transaction set. A data segment is a group of related data elements. For example, one data segment might include the data elements related to the buyer's address, and another data segment might include the data elements related to the supplier's address. The data dictionary defines the exact content of the data elements

Table 1 Partial list of X12 Transaction Sets.

Transaction	Code
Invoice	810
Operating expense statement	819
Payment order/remittance advice	820
Customer account analysis	822
Lockbox	823
Planning schedule with release capability	830
Price/sales catalog	832
Request for quotation	840
Response to request for quotation	843
Product transfer account adjustment	844
Price authorization acknowledgment/status	845
Inventory advice	846
Response to product transfer account adjustment	849
Purchase order	850
Purchase order acknowledgment	855
Shipping notice/manifest	856
Purchase order change request	860
Receiving advice	861
Shipping Schedule	862
Report of test results	863
Purchase order change request acknowledgment	865
Product transfer and resale	867
Order status inquiry	869
Order status report	870
Functional acknowledgment	997

that comprise transaction sets. For example, the standard could specify the invoice identifier data element in the invoice transaction set as the unique identifier for an invoice.

The segment directory contains the formats and definitions for the data segments used in creating transaction sets. For example, the invoice transaction set may require an invoice identifier that may be 4 to 10 numeric characters. A loop is a repeating group of segments or segment groups that are part of a transaction. For example, an invoice transaction segment may contain loops consisting of line items for that invoice.

The X12 standard defines transmission control standards for the electronic envelope required to interchange data. Several types of transaction groups destined for the same recipient may be included in a single interchange envelope. Figure 2 illustrates an example of the X12 interchange format. In this example, one interchange envelope includes two shipping notices and one invoice grouped together for transmission to a single customer. There is a single functional header for the shipping notice documents. Each shipping notice has its own transaction set header and transaction set trailer.

Industry Conventions for Use

The X12 transaction set definitions provide users with many potential data elements that can be included in a business document. For example, the Purchase Order (number 850) transaction set defines over 160 segments of data, and each segment can have up to six data elements (dmx.com, 2002). Several industries have developed conventions to specify which segments and which data elements the industry will use in various documents. The automotive industry and airline industry have each developed a set of conventions to specify the standards of use in their respective industries. In lieu of industry conventions, larger companies may dictate to smaller trading partners what conventions to use.

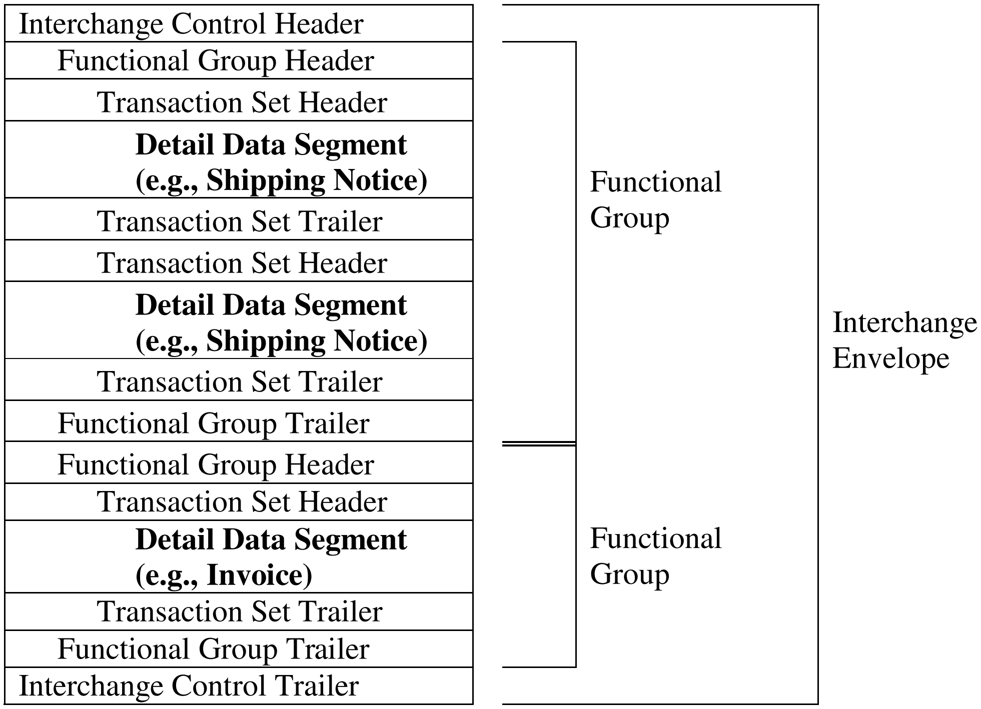


Figure 2: X12 EDI Interchange Format.

The X12 standard has been evolving. The organization responsible for managing EDI standards in the United States is the Data Interchange Standards Association, Inc. (http://www.disa.org). The organization manages both the X12 standards and the EDIFACT standards (Houser, Griffin, & Hage, 1996).

The X12 standard specifies the format and grouping of the data and transaction sets. A data interchange standard is independent of the communications media used to transmit the business documents. A company could choose to use traditional EDI via a VAN, Internet-based EDI, or a proprietary communications network using the X12 standard.

INTERNET-BASED EDI

The Internet provides an inexpensive communications network for performing EDI, so several companies have elected to use the Internet to perform EDI. There are several ways to use the Internet to conduct EDI transactions, but there is currently not a standard approach. Some of the Internet approaches allow organizations to bypass more costly VAN networks. VAN providers have responded by lowering EDI fees and by offering Internet-based services. Companies are using each of the following approaches: exchanging files of transactions using the file transfer protocol (FTP) (dmx.com, 2002); employing a VAN or related service (dmx.com); exchanging transactions via e-mail using multipurpose Internet mail extensions (MIME) (Chan, 2002); or simply using the Web as a means of entering transactions (Bell Helicopter, 2002).

Using the FTP approach to Internet-based EDI may require additional software and custom programming, but could use compression techniques to reduce bandwidth requirements (dmx.com). This usually requires creating a user id and password for each trading partner, and agreement on naming files and directories for messages (Houser et al., 1996).

VAN providers are not standing idly by while EDI moves to the Internet. Many VAN providers are emphasizing their other services while allowing their customers to use the Internet as the communications network for EDI transactions. VAN services can include encryption to secure transactions, compression to speed communications, and software services to ensure the reliable completion of the EDI process (dmx.com).

MIME is an extension of the Internet e-mail protocol that allows the Internet to carry many types of messages via e-mail (Chan, 2002). The extension supports both the X12 and EDIFACT standards for EDI, but specifies how to translate them into a MIME format that a company can send via the Internet (Chan, 2002). The EDI transactions travel via the Internet's e-mail system and the receiving end converts them from the MIME format back to the EDI standard. A company could set up various e-mail destinations for different types of transactions or based on the company submitting the transactions.

Bell Helicopter employs an Internet-based EDI technique that it classifies as Web-based EDI. Using this approach, a trading partner enters the transaction information into an online form and transmits the transaction to Bell Helicopter (2002). The approach provides Bell Helicopter with the advantages of capturing the transaction electronically. It is inexpensive to Bell Helicopter's trading partners because the trading partner only needs a Web browser connected to an Internet service provider (ISP). This approach does not require a significant level of technical expertise on the part of the trading partners. Using this approach, the trading partner does not benefit from capturing the transaction data electronically into its business system.

Many EDI software providers offer software that companies can use for Web-based EDI (Messmer, 2000). For example, one vendor's software includes a server component that includes business forms for entering transactions and another component that a trading partner can download to a Web browser. The company with the Web server software can define the forms for its trading partners to use. The trading partners can fill in the appropriate forms via their Web browsers and submit the transactions to the company for processing along with other EDI transactions.

Extensible markup language, XML, offers the potential for Internet-based EDI. XML is a language for defining documents containing data. It has advantages over HTML because it separates the structure of the data from the content and display of the data. One could define a document in XML and then use that document definition to transmit data across the Internet. It would be possible to define XML documents that contain the same data elements as X12 documents. An extension to XML, electronic business XML (ebXML) is currently under development.

Advantages and Disadvantages of Internet-Based EDI

The two major advantage of Internet-based EDI are the reduced communications costs and the widespread availability of the Internet (Chan, 2002). VAN networks typically charge fees based on usage levels whereas Internet access is typically a flat fee. Companies that have Internet connections can use the Internet for EDI and other applications.

The Web-based EDI employed by Bell Helicopter provides a way for larger trading partners to obtain the benefits of EDI without requiring a significant investment of money or technical expertise from its small trading partners. The larger organization can still process transactions in an automated way through its EDI interface.

The three primary disadvantages of Internet-based EDI are the lack of standards, security, and bandwidth and reliability of the Internet. As discussed previously, there is not a single common approach to Internet-based EDI. Because the Internet is a public network, it is inherently less secure than private networks. Lack of security of EDI transmissions is a disadvantage of Internet-based EDI. Companies can use several of the Internet approaches securely if they employ encryption techniques. The Web-based approach used by Bell Helicopter could use a secure http connection to ensure the security of EDI transactions. When using a VAN network, the VAN provider ensures adequate bandwidth and network availability (Chan, 2002). Because the Internet is a public network, there is no organization that guarantees its bandwidth and availability.

Some VAN providers are encouraging customers to use the Internet to exchange EDI transactions with the assurance that the VAN network is also available (dmx.com).

Despite the potential drawbacks of the Internet, companies are increasingly using the Internet for EDI. As more companies use the Internet, one or several standard approaches will emerge. The security of the Internet is increasing and there are already technologies available for securing transactions.

EDI IMPLEMENTATION ISSUES

Organizations that implement EDI must address the following issues: trading partner relationships, selection of one or more EDI approaches, development of an EDI infrastructure, system and business process changes, and user training. A company's relationships with its trading partners affect the ease of implementation of EDI. Some companies establish trading partner agreements that specify practices for EDI. Houser et al. (1996) suggest that these agreements be kept simple or that existing agreements be used for EDI. Industry practices may dictate what arrangements a company needs to make. It is generally easier to add EDI links with companies that are already doing business via EDI than with a company that is not using EDI. Larger companies may be able to coerce smaller trading partners to adopt EDI, but they should be aware of the costs to smaller trading partners.

The EDI process will likely mean a change in the way a company conducts business. For example, if a company is going to receive shipping notices via EDI, can the company integrate electronic documents into a business system without the need to rekey information? A company will need to maintain its current processes until its trading partners are all using EDI. Each time the company adds a partner to the EDI process requires an incremental effort. For example, the company must add the partner to a list of EDI partners, and the company should test each type of transaction with the partner.

The organization needs to plan which documents to exchange via EDI. Each additional document type requires an additional program to extract the data from or load the data to an existing system. For example, if the company decides to submit invoices via EDI, a program in the Billing System needs to extract the invoice information for translation by the EDI translation software.

Organizations that have greater levels of EDI and technical knowledge are able to implement EDI more fully than other organizations (McGowan & Madey, 1998). EDI has both technical and administrative components. Larger organizations may find that providing assistance to smaller trading partners will improve implementation. McGowan and Madey also found that organizations that made training available for EDI attained greater levels of implementation with EDI. The staff responsible for EDI needs to become familiar with the EDI software and tools.

EDI requires a significant organizational commitment. The costs (see following section) of EDI are not trivial. Top management support is necessary to ensure that resources are available for successful implementation. Organizations need an EDI champion, a person who can push the concept and overcome organizational inertia, to move forward with EDI.

Implementation Costs

The costs of EDI include the cost of data extraction and load programs, translation software, and communications costs. In-house programming staff typically develop the programs to extract data from or load data to existing systems.

Most companies purchase the software used for step two (translation to EDI format) and step three (software to manage document transmission) from an EDI software provider.

The communications costs depend on the alternative a company chooses. With traditional EDI, a VAN provider typically charges a flat monthly fee for the service and a usage fee based on the volume of transactions. A company can use Internet-based EDI either with or without the aid of a VAN. The communications costs are typically much less for Internet-based EDI than for traditional EDI.

A proprietary system requires the purchase or development of software to manage the electronic mailboxes. It also requires the hardware and network infrastructure to handle the EDI traffic load. This approach is not feasible for small companies, but large companies may find it cost-effective.

EDI Implementation and Business Strategy

Organizations should also consider how EDI applications will support the business and what type of system is appropriate for the company's competitive situation. A company may choose to implement EDI as a means of obtaining general benefits such as improved transaction efficiency, reduced errors, or lower costs. An organization could also implement EDI as a tactic in support of a specific business strategy such as just-in-time (JIT) inventory handling or as part of a continuous replenishment process (CRP).

Choudhury (1997) suggests there are three types of interorganizational information systems (IOISs), information systems shared by two or more companies: electronic dyads, electronic monopolies, and multilateral IOISs. He identifies two transaction characteristics (demand uncertainty and market variability) that influence the choice of IOIS. Because EDI is a type of IOIS, this model provides insight into an organization's choice of EDI applications. The model asserts that the selection of IOISs depends on the nature of the relationship an organization wants to implement with its suppliers/buyers.

An electronic dyad is a trading relationship in which a buyer purchases a particular product from one of a small set of suppliers. In an electronic monopoly, a buyer has one (exclusive) supplier. This situation could exist due to a market monopoly (only one supplier exists) or by choice (the buyer agrees to a sole-source supplier). A multilateral IOIS is one in which a buyer shops the entire market for each purchase. Purchasing an airline ticket through the Sabre reservation system would be an example of a multilateral IOIS. EDI is not well suited to this type of relationship because an organization addresses an EDI document to a particular trading partner. Further development of Internet-based EDI may permit this form of exchange.

Choudhury's (1997) contention is that electronic monopolies and electronic dyads both permit a greater degree of electronic integration than multilateral IOISs. He suggests that low market variability favors electronic dyads and electronic monopolies, the types of relationships well suited for EDI. For products with low volume uncertainty, the electronic dyad is the favored approach.

Iacovou, Benbasat, and Dexter (1995) make the following recommendations for EDI initiators, organizations that are encouraging small partners to use EDI, based on their study of EDI in small organizations:

Begin by developing a plan that includes small partners, even if they plan to start with large trading partners.

Assess the preparedness of each trading partner by examining resources (information technology expertise and financial ability) available and the degree to which EDI offers perceived benefits.

Provide technical and financial assistance to partners that are not ready for EDI.

Promote the benefits of EDI to the trading partner.

As a last resort, use coercive tactics, but keep in mind that the trading partners will not benefit as much from EDI when they are unmotivated, unprepared, and feel forced to implement it.

Hart and Saunders (1998) offer advice that is consistent with this. Their study found that when EDI initiators used cooperative strategies based on building trust in partner relationships, the partners were more like to have a greater volume of EDI transactions and to support a greater variety of transactions than when the firm used its power to coerce trading partners to use EDI. Both of the studies mentioned here emphasize the importance of dealing with the organizational as well as the technical issues of implementing EDI.

ELECTRONIC COMMERCE APPLICATIONS OF EDI

EDI is an effective technique for electronic commerce. Companies use EDI primarily for business-to-business (B2B) applications, the largest segment of electronic commerce. There are about 40,000 companies using EDI today (Houser et al., 1996), but the industries that lead in the use of EDI include the auto industry, the airline industry, the banking industry, and the credit card industry (Copeland & Hwang, 1997). The President of the United States, in 1993, issued an initiative to streamline the procurement process, and this is the fastest growing area of EDI (Copeland & Hwang, 1997).

One can use several dimensions to classify EDI applications. Previous sections of this chapter considered the standards (X12, EDIFACT, and proprietary) used in the application. One can also describe the different types of EDI applications based on the communications technology (VAN, proprietary network, Internet-based) used. People frequently differentiate information systems based on the business functions or processes the system serves. For example, EDI applications could include procurement, billing, logistics, payment, or other functional areas.

The following example illustrates the possible use of EDI in the procurement process. If a buyer has EDI links with multiple suppliers, the buyer could initiate an electronic Request for Quote (RFQ) to its suppliers. The RFQ would contain the same information as a paper RFQ, including product specifications, quantity required, and date needed. The suppliers who believed they could satisfy the quote would respond to the buyer with Quotations via their EDI links. The supplier would review the Quotations, select one, and initiate a Purchase Order via EDI to its selected supplier. The Purchase Order would include terms, conditions, and the information that would typically appear in a paper purchase order. The supplier would accept the Purchase Order by responding with a Purchase Order Acknowledgment via the EDI link.

BPR and CRP Applications

Organizations can implement EDI to obtain efficiencies such as improved processing speed or reduced human labor costs. However, companies achieve the greatest benefits from EDI when they implement process innovations concurrently with the implementation of EDI. Business process reengineering (BPR) and continuous replenishment process (CRP) are two types of process innovations companies pursue in conjunction with EDI implementation.

BPR involves making significant changes to organizational processes, a type of organizational change that is significantly greater than merely automating a function. For example, one could automate the Accounts Payable function by providing clerks with automated tools to enable them to match an invoice, a product receipt, and the purchase order for the product. Using BPR techniques, one could eliminate the invoice by paying the price on the purchase order for the number of units of product indicated on the product receipt. Successful BPR requires attention to organizational changes in addition to technological changes. Organizations considering the implementation of EDI should also examine the underlying processes involved to see if they should also pursue BPR.

CRP is a process innovation in which the vendor manages a retailer's inventory, replacing the typical approach in which the retailer places an order with the vendor (Lee, Clark, & Tam, 1999). This is a significant change in the inbound logistics for the retailer and a change in the traditional industry approach. Companies can use EDI to support CRP because EDI provides a way for the retailer to report inventory levels to the vendor. The vendor manages the retailer's inventory. Thus, the application of EDI is supporting a radical process change rather than simply automating the ordering process.

Lee et al. (1999) describe the use of EDI to implement a continuous replenishment process between Campbell and 31 grocery retail chains. Campbell managed the retailers' inventories, and this represented a significant change from traditional retail practices (Lee et al., 1999). Although some retail chains had previously used EDI to place orders, the retailers were responsible for managing their own inventories. The study found that the retail chains benefited from CRP and EDI because they achieved a significant increase in their inventory turns and reduced stockouts at the same time. As with other information technology investments, implementation requires attention to organizational as well as technological issues.

Internet-Based EDI Applications

The use of the Internet for EDI applications is rapidly growing. Many organizations are developing new EDI applications using the Internet, but still support traditional forms of EDI. Bell Helicopter (2002) is offering several EDI options to its suppliers as part of its electronic commerce initiative. Bell Helicopter planned a two-phase implementation. In the first phase, all suppliers would be set up to conduct business electronically. Phase two would involve the rollout of 10 types of EDI transactions. The company planned to have 1,000 suppliers active on EDI within a year. Suppliers had the option of using Internet-based EDI over the Web. For the Internet-based EDI, Bell Helicopter required suppliers to have an Internet connection and Web browser. The supplier could then enter transaction information via the Web browser for transmission to Bell Helicopter.

Wal-Mart began moving to Internet-based EDI in 1998 as part of a strategic and tactical initiative (Frook, 1998). Wal-Mart used packaged software to provide EDI services to smaller companies. The system reduced VAN costs for Wal-Mart and fit with the company's strategy to use more local suppliers. Some smaller suppliers had been unable or unwilling to absorb the expense of using VAN services. The system allows Wal-Mart's trading partners to register and download the required software through a self-service option. The system will permit small companies to automate transactions via EDI and allow verification of transmissions. Wal-Mart plans to develop a global EDI network. It has contracted with the software provider to develop foreign language versions.

Dal-Tile International Inc. implemented an Internet-based EDI system in 2001 to satisfy one of its major customers, Lowe's Companies Inc. (Greenemeier, 2001). The new system has resulted in closer coupling between the two companies. Dal-Tile implemented the EDI system to exchange documents including purchase orders and invoices, and the companies now use the system to share sales information, allowing Dal-Tile to better meet Lowe's needs. Dal-Tile did not previously obtain the sales information via EDI because of the costs associated with such a high volume of data. Dal-Tile is migrating from its traditional VAN EDI system and expects to recover its costs in the year it will take to complete the migration.

J. B. Hunt is a $2 billion trucking company that is moving its EDI away from the VAN approach to the Internet (Karpinski, 2001). J. B. Hunt links to 250 suppliers via EDI and has used traditional EDI standards. The company plans to use new formats using XML over the Internet. J. B. Hunt is concerned about the increased network management it will have to assume and is apprehensive about the change to new formats. The company will begin the conversion by moving its 12 largest customers, representing 30% of its revenues, to ensure quick payback of its $200,000 investment. Some of its customers want to stick with the managed services provided by their VAN providers. The company also plans to add delivery-tracking data to the Internet-based EDI, with an expected savings of $12,000 per week. J. B. Hunt expects the Internet-based approach to reduce the effort required to add a new partner to EDI from the current 40 hours to about 8.

An XML/EDI Application

The following is an example of an Internet-based application of XML for EDI for a book ordering application. Figure 3 illustrates how such a system might operate. The book publisher uses XML to receive orders from distributors and orders from small bookshops. A clerk at a small bookshop can access the publisher's order entry system online, and enter the appropriate information into an online form. The information is stored as an XML document and transmitted to the publisher's order processing system. The book distributor uses an automated inventory management system to automatically generate an order for transmission to the publisher. The distributor generates the order in XML format and transmits it to the publisher. The publisher can process the order from the bookshop and the order from the distributor in its order processing system.

The distributor and the publisher must use the same format in preparing the order in XML format. Figure 4 shows the XML code for a simple order document. By using the same format, the publisher will be able to read and process the order. XML requires a document type definition, but offers the advantage that one can determine whether a document is valid. For example, if the order document sent by the distributor did not include an ISBN, the publisher could not process the order.

One of the current impediments to using XML for EDI is that there are few industry standards. Various industry groups are working on developing standards. EDI standards can serve as a starting point for developing XML standards.

These applications demonstrate the value of Internet-based EDI. Large companies are able to obtain significant savings by avoiding volume-based VAN charges. The Internet-based approach makes sense for a variety of industries. It also permits small companies to meet the needs of their larger trading partners without the expense and technical expertise required for traditional EDI systems.

BENEFITS OF EDI

Organizations ascribe many benefits to the implementation of EDI. Among the benefits attributed to EDI are lower transaction costs, higher productivity, reduced cycle times, increased accuracy, reduced shipping discrepancies, and lower inventory costs (Copeland & Hwang, 1997; Raghunathan & Yeh, 2001). EDI can lower costs by eliminating paper processing. If an EDI system were fully integrated with other information systems, it could eliminate steps requiring human intervention, thus increasing productivity. EDI transactions move electronically, so EDI reduces the time needed to complete a transaction. Moving information electronically also improves accuracy because nobody has to rekey data into another system.

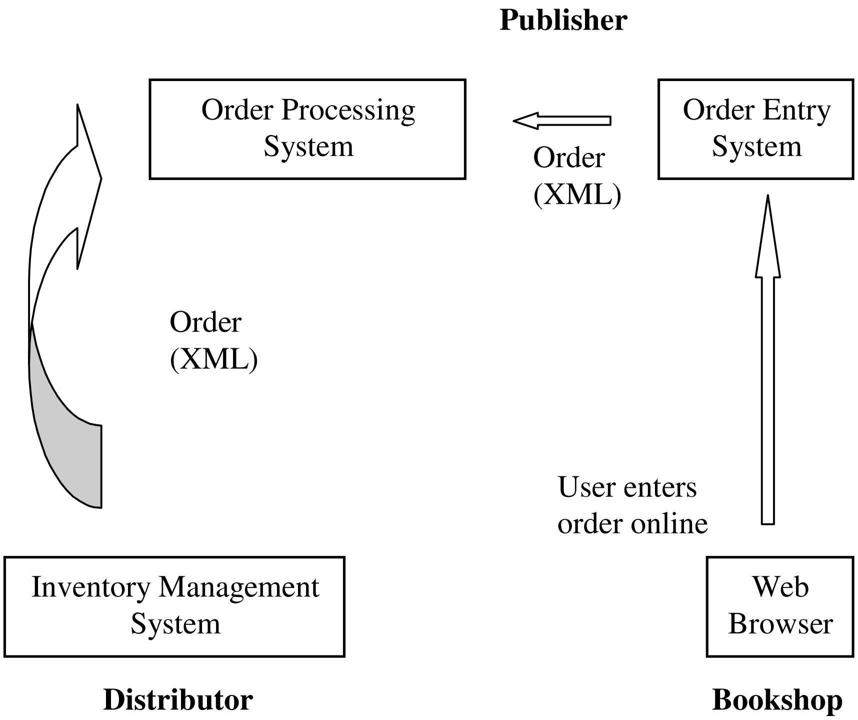


Figure 3: Book order processing using XML/EDI.

It is crucial to note that the degree of benefit, or even whether an organization obtains any benefit, is largely dependent on the specific implementation and application of EDI. If an organization does not have a high degree of automation and the EDI application lacks integration, then it will not realize the full benefit of EDI (Copeland & Hwang, 1997). A company may not obtain these benefits without business reengineering (Lee, Clark, and Tam, 1999). Lee et al. found that it is possible for the EDI initiator and its smaller trading partners to both benefit from EDI through interorganizational reengineering. Their study of the use of EDI by Campbell and 31 grocery retail chains found that the grocery chains reduced stockouts and increased inventory turns by reengineering the supply channel using EDI.

```
<Order>
  <CustomerNumber>10017</CustomerNumber>
  <Quantity>3</Quantity>
  <Description>
    <Title>Beginner’s Guide to XML/EDI</Title>
    <ISBN>123456789</ISBN>
    <Price>22.95</Price>
    <Author>
        <LastName>Doe</LastName>
        <FirstName>Jane</FirstName>
    </Author>
  </Description>
</Order>
```

Figure 4: XML code for a book order.

POTENTIAL DISADVANTAGES OF EDI

EDI has some potential disadvantages, particularly for smaller organizations. Interorganizational information systems projects have higher levels of risk than traditional information systems projects because of their interorganizational nature (Riggins & Mukhopadhyay, 1994). For a company initiating the EDI implementation process, the benefits depend partly on a trading partner’s implementation. The reduced control may lead to higher than expected costs and lower than expected benefits.

Although there are many potential benefits of EDI, in many cases only one trading partner obtains the benefits of implementing EDI. In particular, the trading partner that pushes for EDI is the one that realizes the benefits of EDI (Lee et al., 1999). Even for the larger trading partner, the benefits of EDI may be a long time coming. An organization implementing EDI frequently needs to maintain its traditional business procedures until all its trading partners are EDI capable. It could take years to approach 100% participation by trading partners. In the Campbell CRP case (mentioned in the Applications section), Campbell and its retailers both realized benefits from using EDI-enabled CRP, though some retailers had a learning period of over 6 months (Lee et al., 1999).

Small organizations often adopt EDI only at the insistence of larger trading partners, often their large customers. A large company may coerce its smaller trading partners into adopting EDI, and the smaller organizations may not realize any performance improvements from EDI (Lee et al., 1999).

For smaller organizations, implementing EDI may drive up the cost of doing business without commensurate benefit. The costs of doing EDI usually include translation software, the cost of VAN services, the costs associated with extraction/load programs, and the costs associated with learning the new technology. Small organizations often do not have the volume of transactions needed to make EDI cost-efficient. They act to keep from losing business to a rival, not to obtain a competitive advantage. A small business may not have much of an information infrastructure and may not integrate EDI into its business processes.

To a small business with little EDI integration, EDI means a more expensive way of doing business. A small business might decrease its data accuracy by having to key data into an EDI software package in order to send the document to its trading partner electronically. Small businesses may not possess technical expertise in using the EDI software, and may lose their traditional (paper) audit trails. They may become dependent on the EDI software vendor or VAN for conducting business. The lack of technical expertise may lead to poor security for EDI transactions; unauthorized transactions and fraud may be the result.

Organizations that are smaller than their trading partners may experience a loss of control because they have little influence on whether to use EDI and how the partners will implement EDI (Riggins & Mukhopadhyay, 1994). Smaller organizations may not understand EDI well enough to understand how they will benefit from its use. Organizations initiating EDI should consider investing resources to educate smaller trading partners on the benefits of EDI and providing technical help to smaller trading partners. This may defer the payback for an EDI investment, but may lead to greater levels of integration and better relations among trading partners.

CONCLUSIONS AND THE FUTURE OF EDI

EDI continues to grow. Many companies, particularly larger ones, can benefit from the efficiencies offered by EDI. Companies can use EDI to support JIT and CRP strategies. The benefits of EDI are not as readily accessible to smaller organizations because of the costs and learning required to implement EDI. The shift of EDI to the Internet is leading to reduced costs and greater benefits for small companies as well as large ones.

The brightest options for the growth of EDI are the Internet-based EDI opportunities. The standards committees responsible for EDIFACT and X12 have agreed to develop standards for Internet-based EDI. The most promising development is the work on electronic business XML, ebXML, a standard that extends XML for electronic business. The emerging standard has the potential to reduce costs and simplify EDI connections, particularly for small businesses. Internet-based EDI will provide the benefits of traditional EDI, but with a lower cost structure.

Organizations are likely to attain the greatest benefits when they work cooperatively with trading partners and seek to maximize value in the supply chain. Companies should look for ways to use EDI to support business strategies.

GLOSSARY

DISA (Data Interchange Standards Association) The organization responsible for the development of EDI standards in the United States.

ebXML (electronic business XML) An extension to XML designed to accommodate electronic business.

EDI Electronic data interchange, the computer-to-computer exchange of business transactions in standardized formats.

EDIFACT A set of standard formats for EDI transactions developed by a committee from the United Nations (UN).

Internet-based EDI The use of the Internet as the communications mechanism for electronic data interchange.

IOIS (interorganizational information system) An automated information system that involves information sharing between two organizations.

VAN (value added network) A company that provides a network and various other services for EDI.

X12 A set of standard formats for EDI transactions developed by the American National Standards Institute (ANSI).

XML (extensible markup language) A Web data language designed for representing data structures and data values.

CROSS REFERENCES

See *Business-to-Business Electronic Commerce; Electronic Funds Transfer; Electronic Payment; Extensible Markup Language (XML); Public Networks; Standards and Protocols in Data Communications.*

REFERENCES

Bell Helicopter (2002). *Bell Helicopter's Web-based EDI plan*. Retrieved April 17, 2002, from http://www.bellhelicopter.textron.com / content / eCommerce / edi/ webplan.html

Chan, S. C. (2002). *Introduction to electronic data interchange (EDI)*. Retrieved May 15, 2002, from http://home.hkstar.com/∼alanchan/papers/edi

Choudhury, V. (1997). Strategic choices in the development of interorganizational information systems, *Information Systems Research, 8*(1), 1–24.

Copeland, K. W., & Hwang, C. J. (1997). Electronic data interchange: Concepts and effects. In *INET 97 Proceedings*. Retrieved May 22, 2002, from http://www.isoc.org/inet97/proceedings/c5/c5_1.htm

Dmx.com (2002). The basics of electronic commerce and electronic data interchange. In *Electronic commerce in the public sector: A planning guide*. Retrieved May 17, 2002 from http://www.dmx.com/edibasic.html

Frook, J. E. (1998). Wal-Mart opens its arms to Internet EDI. *InternetWeek.com*. Retrieved June 22, 1998, from http://www.internetweek.com/vertical/retail-48.htm

Greenemeier, L. (2001). Tile manufacturer turns to Internet-based EDI system. Retrieved November 12, 2001, from http://www.informationweek.com/story/IWK20011108S0022

Hart, P. J., & Saunders, C. S. (1998). Emerging electronic partnerships: Antecedents and dimensions of EDI use from the supplier's perspective. *Journal of Management Information Systems, 14*(4), 87–111.

Houser, W., Griffin, J., & Hage, C. (1996). *EDI meets the Internet: Frequently asked questions about electronic data interchange (EDI) on the Internet* (Network Working Group RFC 1865). Retrieved May 23, 2002, from http://www.doclib.org/rfc/rfc1865.html

Iacovou, C. L., Benbasat, I., & Dexter, A. S. (1995). Electronic data interchange and small organizations: Adoption and impact of technology. *MIS Quarterly, 19*(4r), 465–485.

Karpinski, R. (2001). J. B. Hunt's EDI swap-out. *InternetWeek.com*. Retrieved August 15, 2001, from http://www.internetweek.com / transtoday01 / ttoday081501.htm

Lee, H. G., Clark, T., & Tam, K. Y. (1999). Research report: Can EDI benefit adopters? *Information Systems Research, 10*(2), 186–195.

McGowan, M. K., & Madey, G. R. (1998). The influence of organization structure and organizational learning factors on the extent of EDI implementation in U. S. firms. *Information Resources Management Journal, 11*(3), 17–27.

Messmer, E. (2000). Sterling mixes EDI, Web documents. *Network World, 17*(12), 64.

Raghunathan, S., & Yeh, A. B. (2001) Beyond EDI: Impact of continuous replenishment program (CRP) between a manufacturer and its retailers. *Information Systems Research, 12*(4), 406–419.

Riggins, R. J., & Mukhopadhyay, T. (1994). Interdependent benefits from interorganizational systems: Opportunities for business partner reengineering. *Journal of Management Information Systems, 11*(2), 37–57.

Electronic Funds Transfer

Roger Gate, *IBM United Kingdom Ltd., United Kingdom*
Alec Nacamuli, *IBM United Kingdom Ltd., United Kingdom*

PAYMENTS IN THE CLASSICAL WORLD

Payment Systems

Payment systems, which guarantee the safe and efficient transfer of value between counterparties, are an essential requirement for trade and therefore a significant component of the economy's infrastructure.

Excluding cash (fiduciary money), payment systems are a set of mechanisms for the transfer of value among agents, requiring

- An instrument: e.g., check, card, electronic transfer;
- A transfer mechanism and standards: communication channels and messages;
- An audit trail of all transactions submitted and processed; and
- A legal framework and a set of procedures and rules to guarantee settlement finality, or the irrevocable discharge of obligation.

Electronic funds transfer grew up through the need to process higher volumes of transactions faster and more cost-effectively. This drove the need to exchange high volumes of data, initially between financial institutions but increasingly between banks and their customers. The need to remove the inefficiencies of handling cash and checks has driven these volume increases, relentlessly enhanced by the increasing complexity of people's lives. In Europe, payments for utility services and insurance premiums have been a big driver as both businesses and consumers saw the advantages of enhancing their cash flow and budgeting through a move to monthly rather than annual or quarterly payments.

It is essential to distinguish between the information relevant to the payment and the actual transfer of value to achieve final settlement. In mature economies, the central bank always assumes responsibility for final settlement. Commercial banks hold accounts at the central bank and settlement finality (including the legal discharge of the payment obligation) takes place when the account of the payer's bank is debited by the required amount, which is credited to the beneficiary's bank, which will credit his or her account.

When large numbers of payments are involved, they are concentrated in a clearing house, which receives batches of payments for each bank, sorts them by beneficiaries' banks, sends these banks details of the payments for their customers, and calculates net positions, which are settled at the central bank. Bilateral or multilateral netting facilitates settlement by reducing a large number of obligations or positions to a much smaller number directly related to the number of participants.

In these net settlement systems, final settlement only takes place at the end of the day when all payments have been received and net positions have been calculated and transmitted to the central bank for settlement (see Figure 1). Should a receiving bank have acted on the information received and used the funds before settlement, it would have exposed itself to the risk (known as intraday risk) that the payer's bank might not be able to settle its dues at the end of the day.

For this reason, most countries now operate real time gross settlement (RTGS) payment systems for large amounts (typically above $1 million.). The payer's bank sends details of the payment directly to the central bank, which debits his account, credits the beneficiary's bank in real time, and then informs it of the irrevocable receipt of funds. These RTGS payment systems will normally account for less than 5% of the total number of payments within a country but will typically represent over 95% of the monetary value exchanged. Examples of payments exchanged over RTGS systems would include very large corporate payments and the settlement of foreign exchange and securities trades between financial institutions.

The bulk of low-value payments (e.g., salaries, pensions, and payments for utility bills, insurance premiums, and commercial invoices) are handled by automated clearing houses (ACHs). ACHs operate on a net settlement basis but the relatively low amounts involved do not pose an unacceptable risk; they will normally handle two types of electronic payments:

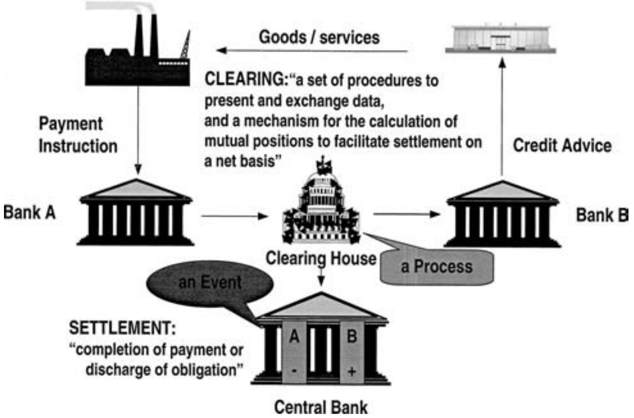

Goods / services
CLEARING:"a set of procedures to present and exchange data, and a mechanism for the calculation of mutual positions to facilitate settlement on a net basis"
Payment Instruction
Credit Advice
Bank A
Bank B
Clearing House
a Process
an Event
SETTLEMENT: "completion of payment or discharge of obligation"
A
B
-
+
Central Bank

Figure 1: Clearing and settlement chain.

- Credit transfers: payers instruct their banks to debit their accounts and transfer the funds to a beneficiary's bank; these include standing orders (STO), whereby the bank is instructed to pay the same amount at regular time intervals.
- Direct debits: payers authorize the beneficiary to draw funds from their accounts: these are normally used for regular payments of variable amounts, e.g., utility or telephone bills and mortgage repayments.

In most countries, corporate customers can submit payment batches (e.g., payroll) directly to an ACH, which will process them after credit authorization from the account-holding bank.

Table 1 illustrates as an example the relative numbers and values of payments handled by the U.S. and U.K. payment systems. In the U.S., the RTGS system is Fedwire and CHIPS is a large-value bank-owned real-time final settlement payments system. In the U.K., CHAPS is the RTGS system and BACS the ACH.

These RTGS systems and ACHs operate nationally for the home currency. An interesting case arose in 1999 when 12 countries of the European Union (EU) joined the European Monetary Union and created a single currency, the euro (€, EUR). A net settlement system in ECUs (the predecessor to the euro) was already in existence and was upgraded by its owner, the EBA (Euro Banking Association), to become Euro1. The European Central Bank (ECB), however, required the establishment of a cross-border RTGS system. Rather than creating a centralized new system, the ECB and the Central Banks wisely decided to interlink the existing national RTGS systems into a decentralized system known as TARGET (Trans-European Automated Real-Time Gross Settlement Express Transfer). Under pressure from the EU and the ECB, systems are currently under development for an ACH-type system for low-value transfers in euros across the EMU countries.

The systems described above relate to payments with the same currency. The explosion of international trade requires, however, the establishment of systems enabling an entity holding its bank accounts in one currency to transfer funds in a different currency. Fundamentally, currencies do not "travel." With the exception of the relatively minute amounts of foreign notes and coins required by tourists, currencies are held and settled in their home countries. International payments are normally settled through a set of procedures known as correspondent banking (see Figure 2).

Table 1 Payments in the U.S. and the U.K. (2001)

U.S.	Daily average value million ($)	% of total value	Daily average volume	% of total volume	Average payment value ($)
Fedwire (including Securities Transfers)	2,561,000	65.5%	513,000	0.7%	4,992,000
CHIPS	1,242,000	31.7%	241,000	0.3%	5,154,000
ACH	60,700	1.6%	21,300,000	27.5%	2,849
Checks	48,100	1.2%	55,335,000	71.5%	1151

U.K.	Daily average value million (£)	% of total value	Daily average volume	% of total volume	Average payment value (£)
CHAPS	333,000	96.1%	105,000	0.5%	3,171,430
BACS (ACH):	8,300	2.4%	13,600,000	66.6%	610
Checks	5,200	1.5%	6,700,000	32.9%	776

Source: Derived from data supplied by Federal Reserve Bank of New York (U.S.) and APACS Annual Review 2001 (U.K.)

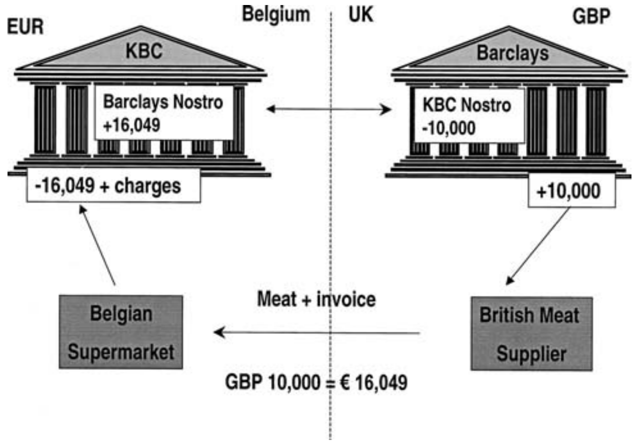


Figure 2: Correspondent banking.

Pairs of banks in each country (correspondents) open "Nostro accounts" for each other in their respective currencies. At the time of requesting the payment in a foreign currency, payers are quoted an exchange rate by their bank and the converted amount in the local currency is debited from their account and credited to the Nostro account of the correspondent bank. Payment information is sent to the correspondent bank, which debits the Nostro account of the payer's bank and credits the beneficiaries in their currency. At the end of the day, each bank sends to its correspondent a statement of the Nostro account for reconciliation.

International payments were amongst the last banking procedures to be automated, owing to differences in languages and national standards. In 1974, a consortium of banks created SWIFT (Society for Worldwide Interbank Financial Telecommunication), which has defined standards for practically every single international financial operation (payments, foreign exchange, securities trading, trade finance, etc). SWIFT operates a worldwide network that in 2001 exchanged daily, on average, 6 million transactions among 7,300 financial institutions in 196 countries, for a total value of $6 trillion. SWIFT also provides the network infrastructure for major international (e.g., TARGET, Euro1) and domestic payment systems.

Figure 3 shows the overall structure of current payment systems.

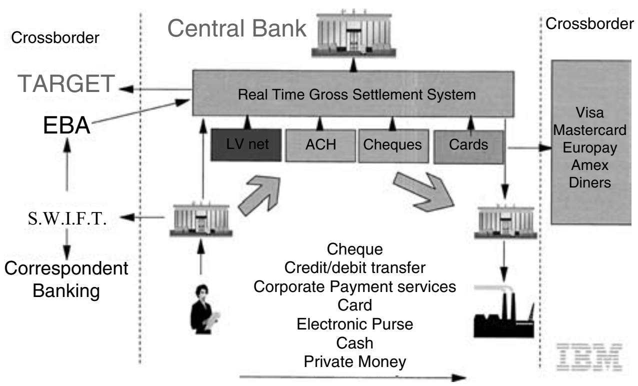


Figure 3: Overall structure of payment systems.

EDI (electronic data interchange) refers to the exchange of structured information related to business transactions. It is common to distinguish between

- Logistics EDI, which relates to the ordering, delivery, invoicing, receipt, acceptance, and possible return of goods if not satisfactory; and
- Financial EDI, which relates to the subsequent payment for these goods.

Standards are defined by various EDIFACT (electronic data interchange for finance, administration, commerce, and trade) Standardization Boards. Under the auspices of the United Nations Centre for Trade Facilitation and Electronic Business (UN/CEFACT), they define syntax and messages for specific business purposes across various economic sectors (automotive, chemical, electronics, finance, etc). Payment instructions would normally include remittance information to ease reconciliation. This remittance information can be extremely large and important when, for instance, the amount paid is less than originally invoiced because part of the consignment is being returned as a result of failing to meet agreed quality criteria. In the U.S. the American National Standards Institute (ANSI) also developed a standard for business-to-business EDI, ANSI X12. This is migrating toward the UN/EDIFACT standard and ANSI is also heavily involved in the development of an XML standard for this area, ebXML. This latter standard is sponsored by two organizations, UN/CEFACT and OASIS, a not-for-profit, global consortium that drives the development, convergence, and adoption of e-business standards. The ebXML mission is to provide an open XML-based infrastructure enabling the global use of electronic business information in an interoperable, secure, and consistent manner by all parties.

Services to Customers

Services to the retail sector are well known: direct payment of salaries into employees' accounts is now widespread and even compulsory in certain countries. Individuals pay for goods and services with checks, cards, and electronic transfers: credits, standing orders, and direct debits. Transaction details appear on statements sent for reconciliation at regular intervals. Most banks now offer Internet banking services enabling customers to initiate payments and receive balances and statements. Payments are not generally perceived as an element of competitive differentiation amongst banks in the retail sector, with the exception of credit cards. In several countries, however, regulators are intervening to force banks to improve execution times and reduce fees for payments. Specifically, the EU has introduced a regulation specifying that cross-border transfers within the Euro zone should not be charged more than domestic transfers.

Competition is more intense in the corporate sector. Large companies, seeking to optimize utilization of their funds throughout the day, require real-time notification of large amounts upon receipt. Banks have developed services enabling corporate customers to maximize the interest earned on their end-of-day balances by aggregating them. This is achieved either by transferring all funds into one concentration account (sweeping) or by offsetting the balances (pooling). Execution deadlines of payments and fees are defined by service level agreements (SLAs) negotiated between the banks and the large corporations. Most banks offer online treasury terminals to their corporate customers for payment initiation, balance and statement enquiries, reconciliation, and other financial services such as foreign exchange. These are gradually being converted to corporate banking portals.

REQUIREMENTS OF E-COMMERCE

E-commerce, although not always living up to expectations, is nevertheless changing purchasing and trading patterns.

Retail Sector

Internet shopping is becoming widely accepted, particularly for travel, groceries, books, CDs and DVDs, and electronics. Customers shopping on the Internet need to pay for goods and services purchased. Utilities and corporations are also moving to electronic bill presentment and payment (EBPP). Credit and debit cards are currently the most popular payment instrument in the B2C (business to consumer) sector, in spite of security concerns.

Cards are, however, not economical for very small payments (less than $10) known as “micropayments,” required for news, games, on-line gambling, etc. Cards are also not suited for P2P (person to person) payments required, for instance, for goods purchased between individuals through Internet auction sites.

Corporate Sector

Trading between corporations on the Internet is increasing. E-markets, where purchasers would seek goods and/or services, negotiate terms, and purchase, have not witnessed the anticipated growth but are viable in certain segments. On the other hand, the Internet has strongly contributed to facilitating commerce between corporations trading across integrated supply chains.

Looking at conventional EFT systems, ACHs are by definition cheap but require a settlement cycle of 2–3 days. RTGSs provide instant finality but at a substantial cost (typically over $20 for a retail customer). Crossborder transfers are also expensive (typically over $15 for a retail customer) and execution frequently exceeds 4 days.

New XML-based industry standards are under development and existing ones are being converted. E-commerce clearly therefore requires faster, cheaper, and above all secure electronic funds transfer systems.

SECURITY IN ELECTRONIC FUNDS TRANSFER FOR E-COMMERCE

History and Background

The changing landscape for EFT has created a number of security challenges that need to be addressed.

In the retail domain growing volume and the proliferation of new Internet payment methods linked to e-mail or prepayment schemes, some of which shift the security issues away from the financial institutions to new players, can alter the underlying security for the user.

In the business-to-business domain payments are becoming more and more embedded in the exchange of data that encompass the complete trade chain. This has brought new routes into being through which payments are initiated and therefore generated the need to create trust infrastructures that can handle these.

Also, as payment services developed, they moved from concentrating the security issues around volume and value to having to take account of the wider range of network channels that had developed, which brought with them new opportunities but also risks.

The methods and tools by which security has been achieved have therefore developed over the past three decades, starting with paper-based systems ranging from simple letters attached to magnetic tapes, and then on to the use of once-only bar code labels stuck to the covers of magnetic media such as floppy disks. As transaction data began to be transferred over networks, initially closed extranets and latterly open networks such as the Internet, new tools such as challenge and response devices that create one-time passwords, message authentication codes to ensure message integrity, and the use of cryptography for both confidentiality and nonrepudiation have been adopted.

Common tools used to deliver security through the development of these services are as follows:

Bar Code Labels

When data is transferred on magnetic media, the automated clearing house (ACH) or bank needs to know that an approved person within the correct organization has sent them. To achieve this that person would be sent a pad of once-only sticky labels, each of which was a once-only password in the form of a printed bar code that could be read by the receiving company. Additional security was managed by ensuring that the labels were kept under lock and key; the data were prepared and transferred onto the media by separate staff and once labeled the media were placed in a tamperproof carrier and sent by courier to the processing center. Risks were also mitigated by the fact that often an input report would be received and checked by a further person within the company prior to the payments actually being settled. These reports would not typically have all items shown, only summary totals, but could include a small number of items, for example those above a certain limit or random entries, which could be checked against the original input data.

It should be noted that this method does not enhance confidentiality or integrity.

Once-Only Password Calculators

As data transmission moved to networks (typically these started on closed user group X25 protocol networks, which were built to internationally recognized standards from the International Telecommunication Union (ITU), and not as open as the current Internet), there was a need to move the password to being a piece of data in itself that was not available for reuse if it was captured during the transmission. This was achieved by issuing customers with a small calculator into which a user could enter a PIN (personal identification number) and then additional data such as the date or payment file value and get back a once-only password in the form of a number that could be added to the header of the data transmission. These numbers were created in a sequence that was synchronized with the host system so that the passwords could be recognized and checked as they were used. These are also known as challenge and response devices.

IDs and Passwords

As e-banking systems based on dumb terminals and later client-based PC software began to allow payments, the logon to these systems was based on IDs and passwords. Frequently these systems asked the user to re-input the password when a specific payment or group of payments was being authorized. However, to reduce the risk in these systems, they frequently used two further tools to ensure authentication and authorization.

First, they frequently limited users to making payments against a predetermined library of beneficiaries. Each beneficiary's details, in particular the account number and bank identifier, are entered into a database by the bank against signed authorizations. This means that the user can only make payments to a subset of beneficiary accounts and does not have the ability to set up new accounts and therefore make a payment to his or her own or a collaborator's account.

Second, the user can use a message authentication code. In the simplest form, certain parts of the payment message, e.g., all or part of the amount, part of the date, and the beneficiary details reference number, known to both the sender and receiver, are added up in a certain way to create a check sum. More complex message authentication codes are created using algorithms to create a check sum made up from the whole message. In either case these calculations can be repeated by the host system and the results compared to ensure that there has been no change to the underlying message and therefore its integrity has not been compromised.

Symmetric Keys

As the demand for PC-based banking systems gathered strength, banks needed to offer free-format payment systems and thus stronger authentication and integrity checking became a requirement. Initially this was delivered using symmetric keys stored either on the PC or on a token such as a swipe card or chip card. Symmetric keys are very good for signing payments and also encrypting data for confidentiality. Messages signed or encrypted in this way can only be checked or decrypted using a copy of the same key, meaning that this had to be carefully protected by each of the parties concerned. This also meant that both parties e.g., the bank and the customer, know the key, so that it is difficult to prove nonrepudiation; i.e., either party could have signed and initiated the message.

Public Key Cryptography

Here there are actually two keys that are inextricably linked. The first is a private key that can be used to encrypt data or sign a message and the second, a public key, is the only key that can decrypt the data or check the signature. Note that the private key cannot be used to decrypt the message or check the signature in this case, because of the way in which the mathematics works. The reverse is also true if the public key is used to encrypt data, in that only the private key can be used to decrypt it. Through the use of various protocols this technology can be used to give nonrepudiation as well as strong encryption of data.

Table 2 summarizes the security characteristics of each of the methods discussed.

A final point that must not be forgotten is that most queries relating to any messages, for particular payments, always happen after the fact, sometimes after many months have passed. Therefore all systems need to have very capable and accessible audit and archive systems that can ensure that queries can be quickly and effectively sorted out. Entries to these audit systems should also be time-stamped using a certified method so that there is a

Table 2 Security methods

Security Method	Message Integrity	Message Authentication	Confidentiality	Nonrepudiation
Bar code labels	No	Authentication of data file only	No	No
Once only password calculators	No	Authentication of the data file only	No	No
ID and password	No	Authentication of the data file or communications session	No	No
Message authentication code	Yes	Provided by the ID and password	No	No
Symmetric keys	Yes	Yes	Yes if used to encrypt the data	No as both parties know the same key
Public key cryptography	Yes	Yes	Yes	Yes

clear and irrefutable sequence for the transactions that were carried out.

Trust Infrastructures Supporting New Payment Initiation Types

Throughout the development of EFT services individual payment messages and files of payments initiated by the bank customer have always been sent to the bank itself or the automated clearing house direct. In trading terms this has meant that the payment message has been created, signed, and sent to the financial institution by the buyer/payer. Also, except in limited EDI message types, the payment has often been divorced from the underlying information or trade cycle. At the same time banks have seen their traditional revenue streams decline as payments have been commoditized, reducing the prices they can charge, and new players have entered the market, taking advantage of new lower cost technologies.

Many banks therefore have recognized the need to develop services that enhance their position in the trade cycle and potentially bring them additional business opportunities. To do this a new trust model needed to be developed to allow payments and messages initiated by one bank's customers to be checked and trusted by another bank and its customers. This has been achieved by linking the development of individual public key infrastructures at banks across the world to a single trust scheme called Identrus. Other schemes such as the Global Trust Authority (GTA) have come to market, but at the time of writing Identrus is considered the most mature.

At the time of writing Identrus included 54 major banks from around the world and was actively recruiting additional members. Each bank has signed agreements that link it to the scheme and is either issuing or preparing to issue digital certificates that confirm the identity of the subscribing partner in such a way that they can be checked with certainty by any of the other banks on behalf of themselves or their customers.

Thus Identrus provides a global framework for the provision of certificate authority services, enabling financial institutions to extend their full range of services onto the Internet and become trusted third parties for e-commerce transactions. Through this they can develop new lines of businessas the Internet becomes a preferred transaction medium and their customers will have the ability to leverage a single bilateral relationship with a financial institution for all e-commerce dealings. At the same time this offers businesses and financial institutions a way to proactively manage the risk associated with e-commerce.

This form of trust has currently given rise to two new payment initiation schemes: Project Eleanor from Identrus and SWIFT e-Payments Plus from SWIFT (it should be made clear at this point that the actual settlement still occurs using the current bank settlement systems: SWIFT, CHAPS, ACH, etc.). They are both broadly similar in that they offer the ability for a payment message to be sent direct to the bank by the buyer or via the seller and the seller's bank. A range of payment types are offered ranging from a straight payment order to payments that are guaranteed by the buyer's bank or where a number of conditions are attached to payment that have to be completed by the seller and confirmed by a third party.

The first of these is a payment order, which is a revocable, unconditional electronic instruction from the buyer requesting the buyer's bank to execute a credit payment to the seller on a specific date for a specified amount. It is revocable in that the buyer can revoke the payment order up to the time the buyer's bank executes it. It is unconditional in that the seller is not required to show it has met its obligations in the underlying contract in order to receive payment.

The second type of instruction is a payment obligation, which is an irrevocable, unconditional undertaking of the buyer to pay a seller, or holder of the obligation, on a specific date for a specified amount at the buyer's bank. The payment obligation is irrevocable in that the buyer cannot revoke the obligation once it has been issued to the seller without the assent of the seller. The reimbursement arrangement for the buyer's obligation becomes effective on appropriate acknowledgement from the buyer's bank that the instruction has been accepted. The seller will have recourse against the buyer should payment not be made. The undertaking is unconditional in that the seller is not required to demonstrate that it has met its obligation in the underlying contract in order to receive payment. The rights of a payment obligation may be transferred, enabling sale and transfer of holdership in a secondary market. Furthermore, a payment obligation may also have the guarantee of the buyer's bank, in which case it becomes a "certified payment obligation."

Finally, a conditional payment may be a payment order ("conditional payment order") or a payment obligation ("conditional payment obligation") in favor of a named seller, payable at the buyer's bank upon presentation of specified electronic message(s), with a valid digital signature created by specified party(ies), to the buyer's bank, as evidence of fulfillment of preagreed conditions. The party(ies) issuing the electronic message(s) may be one or more of the following: the buyer, the seller, a third party service provider (TPSP), or a trusted service supplier (TSS). Such party(ies) must be agreed upon by at least the buyer, the seller, and the buyer's bank. The resultant payment under a conditional payment order or conditional payment obligation may be "immediate" or "deferred." Furthermore, a conditional payment obligation may have the guarantee of the buyer's bank, in which case it becomes a "certified conditional payment obligation." Table 3 summarizes the different payment types.

If we look at the first of these, a payment order using the Project Eleanor protocol from Identrus, the payment flows in a seller-initiated scenario are illustrated in Figure 4.

In the SWIFT e-Payments Plus model (see Figure 5) the main difference is that the messages are transmitted through the Internet to the SWIFT network, which carries out some of the processes prior to passing the data on to the bank. Further details can be obtained from the SWIFT Web site, http://www.swift.com.

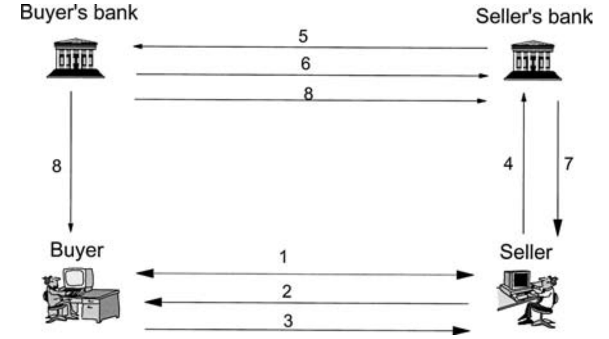
Table 3 Trusted payment types

Product	Revocable	Whose responsibility transferable on	Holdership	Payment condition
Payment order	Yes	Buyer	No	No
Payment obligation	No	Buyer	Yes	No
Certified payment obligation	No	Bank	Yes	No
Conditional payment order	Yes	Buyer	No	Yes
Conditional payment obligation	No	Buyer	Yes	Yes
Certified conditional payment obligation	No	Bank	Yes	Yes

In both cases security is provided by using the private keys generated in the process of registering for an Identrus certificate.

Summary

Security is achieved through authorization, document authentication, and signer authentication. PKI technology such as digital certificates, digital signatures, and encryption can be used to fulfill these requirements. Transaction nonrepudiation is achieved by creating a record that ensures that the agreement represented by the record cannot be disputed; this can also be achieved through the use of digital signatures. To ensure proof that a transaction was correct and properly authorized at the time, a transaction record has to be part of a reliable audit trail for internal (business audits) or external (legal, or chain-of-custody) purposes. To achieve this the transaction record needs to be capable of being archived and reconstructed with its content and context clearly defined.



1) Buyer and seller interact through seller's online system.
2) Seller software presents payment form to buyer.
3) Buyer signs payment and sends it to seller.
4) Seller appends data, signs and sends payment to seller's bank.
5) Seller's bank checks seller's signature, certificate, and that the buyer's bank is an Eleanor member. Seller's bank validates seller account details and appends data. Seller's bank signs and sends payment to buyer's bank.
6) Buyer's bank checks signature and certificate of seller's bank and that the seller's bank is an Eleanor member. Then buyer's signature and certificate are checked. Buyer's bank validates account details. Buyer's bank sends positive Service Response to seller's bank.
7) Seller's bank re-signs the Service Response and sends to seller.
8) On execution of payment, buyer's bank notifies buyer and seller's bank.

Figure 4: Identrus message flows.

B2C AND P2P E-EFT

Retail Payments

As mentioned, most banks offer Internet banking services allowing customers to initiate payments or transfer funds between their accounts.

New entrants are, however, proposing alternative schemes. The most successful is PayPal (now a subsidiary of e-Bay, the leading online auction house) which arose from the need to settle P2P payments between buyers and sellers on Internet auctions.

Customers open accounts with PayPal and register their credit card numbers and/or bank account numbers.

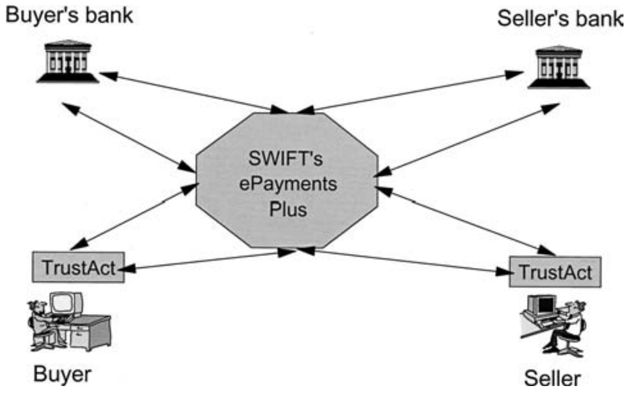


Figure 5: SWIFT e-payments plus.

The Send Money instructions on the PayPal website reads as follows: "Just enter the recipient's email address and the amount you wish to send. You can pay with a credit card or checking account. The recipient gets an email that says "You've Got Cash!" Recipients can then collect their money by clicking a link in the email that takes them to https://www.paypal.com/" (PayPal, n.d.).

Recipients are invited to open PayPal accounts. They can either transfer the funds received to their bank accounts or retain them for further purchases. Fees are significantly lower than those charged by banks and funds are immediately available. The system is viral, as recipients must open aaccounts with PayPal in order to receive the funds.

Mobile telephones appear to be a convenient solution for micropayments. Some solutions have been launched, notably PayBox in Europe and by some Scandinavian banks. Telecomm operators would appear well positioned but lack the credit assessment capabilities to enable them to safely offer services to purchase higher value goods and services. The future probably lies in alliances between banks and telcos. Various standardization bodies have been established, but progress is hampered by the lack of uniformly accepted standards.

ELECTRONIC BILL PRESENTMENT AND PAYMENT (EBPP)

EBPP is defined as the electronic presentment of bills combined with a mechanism to allow the customer to receive and pay the bill electronically. Significant cost savings would accrue to billers: printing, mailing, and in countries such as the U.S. where checks dominate, processing received checks and reconciliation.

Bills can be sent either directly to the customer's e-mail (biller-direct model) or to a consolidator, allowing customers to access all their bills at the same site. Billers now view bills as more than mere financial statements: bills are perceived as opportunities for brand reinforcement, marketing, cross-selling, and customer care. Two consolidation models are possible, with the latter therefore emerging as the favorite:

- Thick consolidation, where the full contents of the bill are presented by the consolidator, and
- Thin consolidation, where the consolidator only presents summary information with a link to the biller's site.

Two roles arise from EBPP:

- Biller service provider (BSP): e-bill extraction, publishing, hosting, and distribution; and
- Customer service provider (CSP): customer enrolment, presentation (thick or thin), customer care, payment.

EBPP has not grown as fast as originally anticipated, as customers are reluctant to pay for a process which today costs them virtually nothing. Critical mass is essential on both sides: billers are seeking to reach the maximum number of customers, whereas customers are interested in receiving the maximum number of bills online. Banks are viewed as natural providers of consolidation services, acting as both BSPs and CSPs on behalf of their corporate and retail customers. In the U.S., an example of a consolidated bill presentment and payment service is offered by CheckFree in partnership with hundreds of billing companies and a range of financial institutions across the U.S. Each time users log onto the service they are automatically directed to their MyCheckFree page, where they can view and pay new e-bills or update existing billing and payment information. Users can then pay their online bills using any ACH-enabled bank account, such as a checking or money market account, or major credit cards where accepted by the biller.

To achieve critical mass, some Scandinavian countries have implemented EBPP systems through their ACHs. The technical architecture of such a system is illustrated in Figure 6. For payment, customers are normally given the choice of paying bills by debiting their bank accounts, direct debit, or charging to credit cards.

EBPP in the B2B sector is sometimes referred to as EIPP (Electronic Invoice Presentment and Payment), but the fundamental principles are the same.

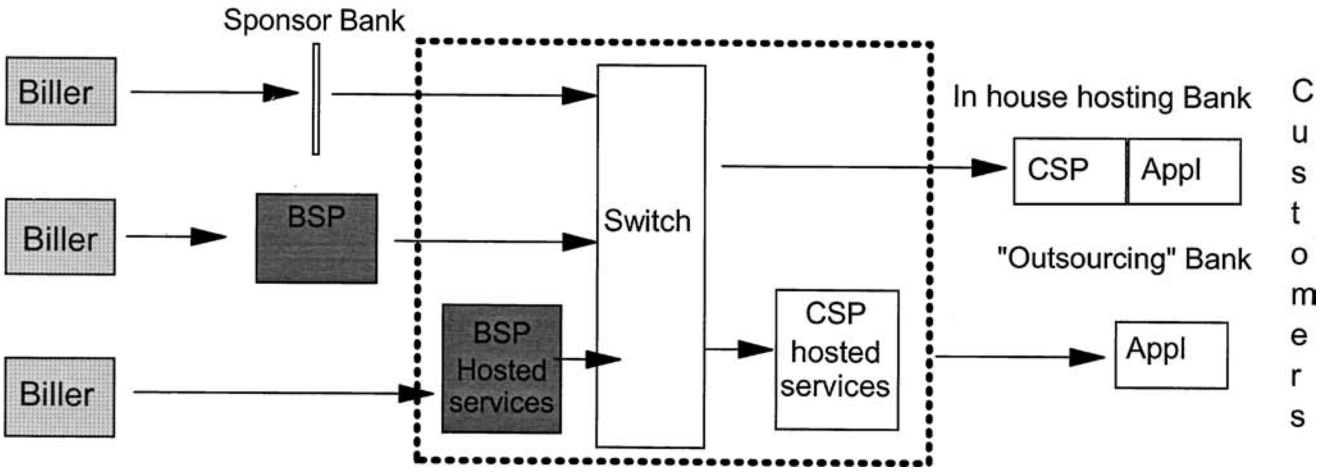


Figure 6: EBPP architecture.

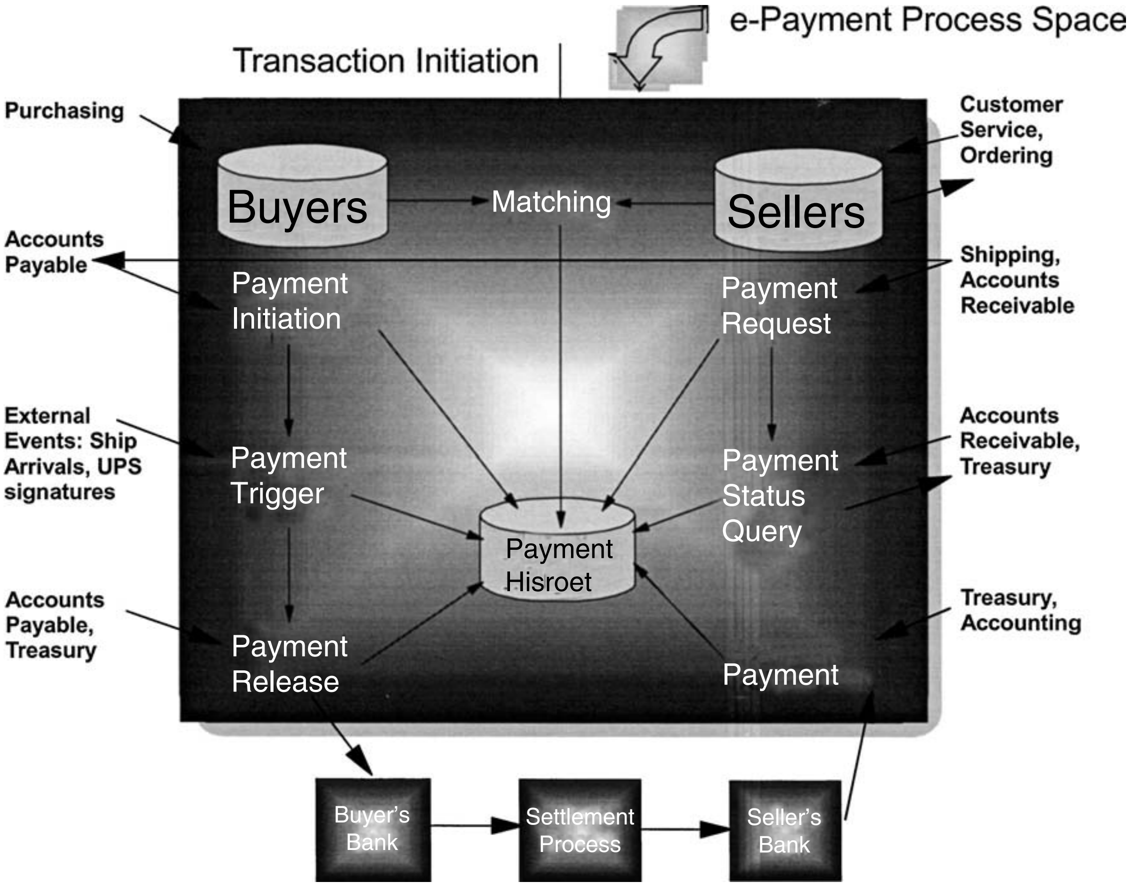


Figure 7: Basic B2B e-payment model.

B2B E-EFT

Requirements for B2B payments are more complex. The payment is one of the final steps of the complete commercial cycle and occurs upon formal acceptance of the goods or services. Purchases orders, stock control, and financials are increasingly being handled by corporate-wide ERP (enterprise resource planning) systems. Any payment solution must therefore be capable of integration with the corporation's logistic and financial processes and systems. Figure 7 illustrates the basic B2B e-payment model and the required features.

The actual transaction could be initiated by the corporation itself, as payment of an invoice presented electronically via EBPP/EIPP, or as the result of a purchase agreed to on an e-market or an integrated supply chain. E-markets and integrated supply chain systems have concentrated so far on seeking partners, negotiating terms, and sometimes logistics. To date, very few have also incorporated financial services, so payments are effected via conventional channels. Some banks are positioning themselves as providers of payment and financial services to e-markets and integrated supply chains.

FUTURE PERSPECTIVES

The systems described above do not bypass the banking system, as payments originate from and end in bank accounts and clearing and settlement take place via the conventional systems (cards or ACH). Possible exceptions could be payments charged to telephone bills, although even those are ultimately paid via the banking system. The danger for the financial institutions lies in potential disintermediation, brand substitution, and loss of customer contact. Customers are now paying through PayPal or their mobile-phone operators and their banks are one level removed.

Particularly in the B2B sector, significant opportunities exist for banks to offer classical banking services, integrated across the commercial value chain, to e-markets or clusters of corporations trading across industry supply chains: market intelligence, information and credit assessment on trading counterparties, credit, trade finance, shipping and customs clearance, escrow, factoring, invoice financing, reconciliation, treasury, etc. (see Figure 8), as well as trust services and hosting cumbersome EDI remittance information. Some banks are already offering such services to reposition themselves at the centers of commercial chains.

The technological and organizational challenge banks face is to integrate processes and IT systems, which are today siloed, into customer value propositions on user-friendly portals.

To conclude, we can predict the following:

- Major structural changes are unlikely to occur under current legal and regulatory environments,

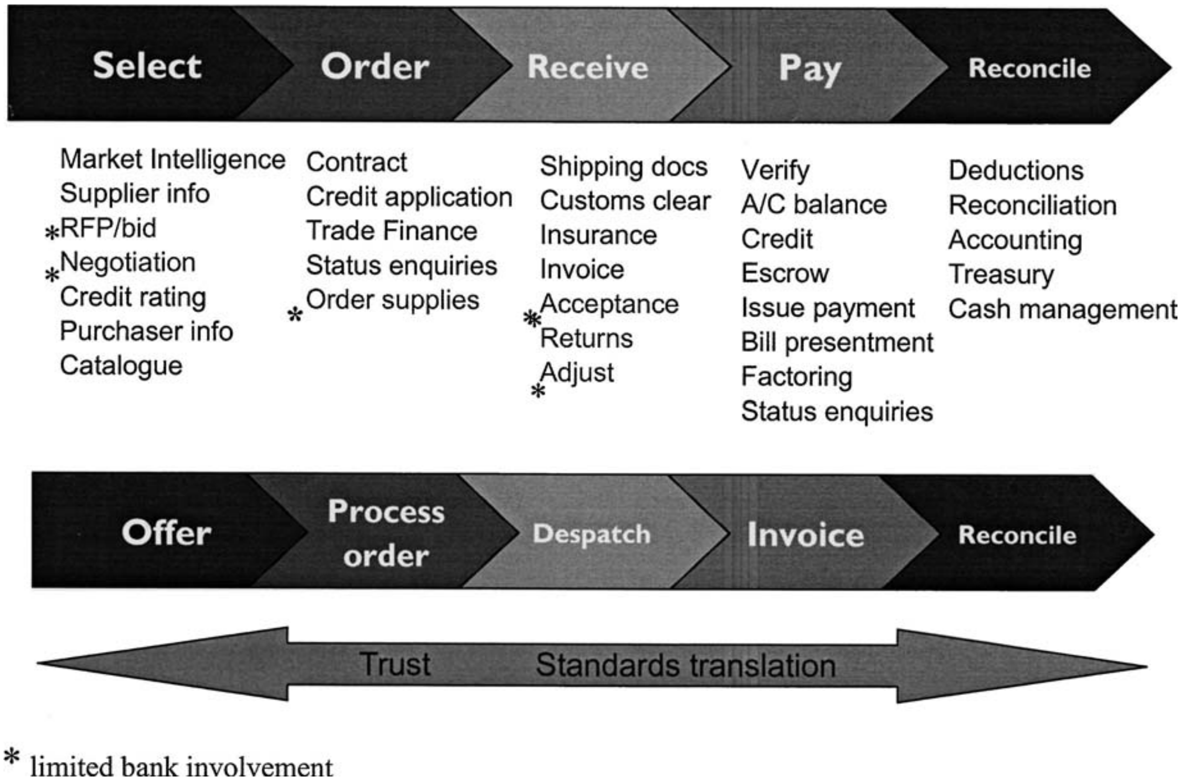


Figure 8: Opportunities for banks throughout commercial value chain.

- Service level agreements, which are commonplace in the corporate sector, will be supplemented by regulation and anticompetitive regulation in the retail sector,
- **Internet payment submission will dominate B2B within 3 years,**
- New payment systems offering speedier finality at lower cost will emerge,
- Competition will increase as banks seek alliances with telecommunication providers and e-markets, offering services to clusters of corporations across industry sectors, and
- Attention will focus on security, fraud detection, anti-money-laundering measures, and business continuity.

GLOSSARY

(Source: Bank for International Settlements Standard Red Book terms)

Automated clearing house (ACH) An electronic clearing system in which payment orders are exchanged among financial institutions, primarily via magnetic media or telecommunication networks, and handled by a data-processing center.

Batch The transmission or processing of payment orders and/or securities transfer instructions as sets at discrete intervals of time.

Correspondent banking An arrangement under which one bank (correspondent) holds deposits owned by other banks (respondents) and provides payment and other services to those respondent banks. Such arrangements may also be known as agency relationships in some domestic contexts. In international banking, balances held for a foreign respondent bank may be used to settle foreign exchange transactions. Reciprocal correspondent banking relationships may involve the use of so-called Nostro and Vostro accounts to settle foreign exchange transactions.

Discharge Release from a legal obligation imposed by contract or law.

Electronic data interchange (EDI) Electronic exchange between commercial entities (in some cases also public administrations), in a standard format, or data relating to a number of message categories, such as orders, invoices, customs documents, remittance advice, and payments. EDI messages are sent through public data transmission networks or banking system channels. Any movement of funds initiated by EDI is reflected in payment instructions flowing through the banking system. EDIFACT, a United Nations body, has established standards for electronic data interchange.

Final settlement Settlement that is irrevocable and unconditional.

Netting An agreed offsetting of positions or obligations by trading partners or participants. The netting reduces a large number of individual positions or obligations to a smaller number of obligations or positions. Netting may take several forms that have varying degrees of legal enforceability in the event of default of one of the parties.

Payment A payer's transfer of a monetary claim on a party acceptable to the payee. Typically, claims take the form of banknotes or deposit balances held at a financial institution or at a central bank.

Payment order (or payment instruction) An order or message requesting the transfer of funds (in the form of a monetary claim on a party) to the order of the payee. The order may relate either to a credit transfer or to a debit transfer.

Payment system A set of instructions, banking procedures, and, typically, interbank funds transfer systems that ensure the circulation of money.

Real-time gross settlement (RTGS) A gross settlement system in which processing or settlement take place in real time (continuously).

Settlement An act that discharges obligations with respect to funds or securities transfers between two or more parties.

S.W.I.F.T. (Society for Worldwide Interbank Financial Telecommunication) A cooperative organization created and owned by banks, which operates a network that facilitates the exchange of payment and other financial messages between financial institutions (including broker-dealers and securities companies) throughout the world. A S.W.I.F.T. payment message is an instruction to transfer funds; the exchange of funds (settlement) subsequently takes place over a payments system or through correspondent banking relationships.

CROSS REFERENCES

See *Business-to-Business (B2B) Electronic Commerce; Consumer-Oriented Electronic Commerce; Electronic Data Interchange (EDI); Electronic Payment.*

REFERENCES

ANSI (n.d.). Retrieved August 2002 from http://www.ansi.org

BIS (Bank for International Settlements) (n.d.). Retrieved August 2002 from http://www.bis.org

EBA (n.d.). Retrieved August 2002 from http://www.abe.org

ebXML (n.d.). Retrieved August 2002 from http://www.ebxml.org

ECB (n.d.). Retrieved August 2002 from http://www.ecb.int

Identrus (n.d.). Retrieved August 2002 from http://www.identrus.com

ITU (n.d.). Retrieved August 2002 from http://www.itu.int

OASIS (n.d.). Retrieved August 2002 from http://www.oasis-open.org

PayPal (n.d.). Retrieved August 2002 from http://www.paypal.com

SWIFT (n.d.). Retrieved August 2002 from http://www.swift.com

UN/CEFACT (n.d.). Retrieved August 2002 from http://www.unece.org/cefact/

Electronic Payment

Donal O'Mahony, *University of Dublin, Ireland*

INTRODUCTION

In 1994, Netscape Corporation released its first Internet browser to an Internet population that consisted of users of approximately 3 million host computers around the world. Before the term *electronic commerce* was coined, companies who had products to sell used the Web as a virtual shop window. Product brochures, specifications, and data sheets, which had hitherto been physically distributed, were put up on Web sites. Many purchasing decisions were made based on this information, but although the details of the deal may have been discussed electronically, the final order, and more importantly the payment, took place by more conventional means. By the end of 1998 (which is remembered by many as the first Christmas holiday season when on-line buying became highly significant), the Internet population had grown to around 43 million hosts worldwide. Just two years later, the number of hosts worldwide had risen to 93 million with on-line retail sales reaching over $10 billion for the holiday season and three times this for the year as a whole. Much of this took place using credit cards as the payment method and in spite of the fact that many consumers are still loath to trust the Internet as a means of effecting payment. Although many candidate systems (O'Mahony, 2001) have been offered to fill the payments gap for business-to-consumer e-commerce, not many have been adopted to any significant degree, and it may take some years before the industry converges on a standard in this area. It is really only since 1999 that companies have started to get very interested in the area of business-to-business e-commerce, and the jury is still out as to what will be the payment method of choice in this environment. In order to gain a greater understanding of what payment methods can be used across the Internet, it is useful to first examine how payment is effected in conventional commerce.

Conventional Payment Methods

The first and most obvious method of transferring value is the use of cash. This evolved from bartering of universally sought-after commodities such as salt or gold into today's system involving notes and coins issued by national governments. In most countries in the world, cash is used for some 80% of day-to-day transactions. It is very versatile in that in can be used for payment to merchants or simply from one individual to another without the need to involve a financial intermediary such as a bank. There is no need for prior trust to be established between the parties, although where this is an issue, the payee will likely examine the notes very carefully before accepting them. Very low-value transactions can be carried out, limited only by the smallest denomination on the coinage; indeed the fact that the average cash transaction in the United States is around $11 demonstrates its main realm of applicability.

It is not without its problems, however—it costs governments a significant amount of money to produce and maintain the national stock of notes and coins and increasingly this process is open to attack from counterfeiters. It can be stolen. Its anonymity makes it attractive to organized crime, and its use in large-value transactions is often associated with tax evasion. Because a given currency is tied to a particular country or region, cash is not very suited to international transactions. In spite of its problems, it is the dominant payment instrument in business-to-consumer and person-to-person low-value transactions.

When it comes to higher value transactions and also for business-to-business payments the use of the check becomes important. A check is simply an instruction to the payer's bank to transfer funds to the payee. It is fundamentally dependent on the presence of a financial intermediary, usually a bank, and unless everyone in the community shares the same bank, there must also be a clearing and settlement mechanism to connect the two banks involved.

The payment instrument can be used to transfer values of any amount, but the involvement of at least one intermediary coupled with a considerable amount of paper processing means that the transaction charges are likely be to around (U.S.) $1. Figure 1 shows the steps that need to be taken to process a check in a conventional clearing system.

The cost of this processing effectively makes checks impractical for very small transactions. Because a check is merely a promise to pay, it depends on a degree of trust being established beforehand between the two parties. Where this is lacking there may be a delay in issuing the goods or the parties may have to resort to some kind of escrow arrangement. Banks often issue check guarantee cards, which indemnify the payee against risks as long

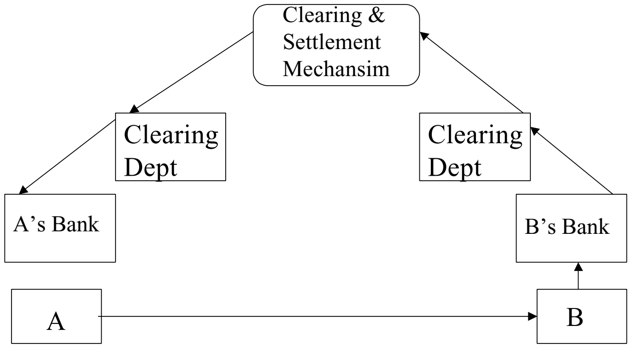


Figure 1: Check clearing process.

as the transaction size is small and some fairly rudimentary security checks are made at the time of the transfer. For larger transactions, a customer's bank will often sign the check itself, converting a simple check into a cashier's check or bank draft. Since the promise to pay is now coming from a much more trusted party, greater risks can be taken on the part of the payee to ensure that the transaction completes.

Typically all parties to a check transaction are in the same country, so no currency issues arise. Where the payer and payee are in different countries, decisions must be made about which currency to use and at what point the conversion is made. In the absence of any globally acceptable clearing and settlement authority, the major banks very often resort to correspondent banking, where special arrangements are made with a bank in the destination country to handle checks issued by them. This necessarily makes the transactions more expensive, and the need to manage risk limits the speed with which the transaction can be carried out.

To some extent, the use of a debit card is similar to an electronic version of writing a check. The source of funds is the same, as is the general flow of information. Typically, though, the debit card does some kind of authorization at the point of sale, verifying at least that the card is not stolen and possibly even verifying that sufficient funds are available.

One of the major problems in check payment is the risk associated with the "returned item" or bounced check. Even though the incidence of checks being returned is very small, the fact that it can happen at all associates high risk with a check and makes it unacceptable for many transactions, particularly where goods are delivered immediately. One way to avoid this risk is to use a credit transfer or Giro payment. Where a check represents a "pull" payment with the paper check pulling funds from the source account through the clearing network into the destination account, the Giro does the reverse. Funds are "pushed" from the source account to the destination account. The credit transfer cannot be initiated unless the funds are available, and this eliminates any risk associated with the payment.

Where the bank details of the payee are known in advance, it is possible to make electronic transfers between bank accounts using so-called automated clearing house (ACH) networks. These organizations grew out of the systems that were developed to process check clearing and are now used by consumers for recurrent payments to utilities in the form of direct debits. They are also used extensively by businesses to pay their regular suppliers and by governments to issue all manner of payments to individuals and corporations. In the United States, the system is operated by the National Electronic Payments Association and most countries in the developed world have similar systems. Indeed, it is quite common to have multiple systems of this sort operating in a single country—some operated by the central bank, and others by consortia of leading banks.

In 1999, the average value of a payment made through the ACH system was approximately $1,500 and settlement is made overnight. Where the value of the transaction is significantly larger, a different class of payment method is typically used, which is referred to as a "wire transfer." One example of this is the FEDWIRE system operated by the Federal Reserve in the United States. This offers the facility to make immediate payments with settlement performed by transferring funds between accounts maintained by the member banks with the Federal Reserve. In 1999, the average value of transactions in the FEDWIRE system was $4.3 million. It is thus used principally for major business-to-business and also business-to-government transfers. When such payments are made internationally, the messages relating to the wire transfers are typically carried on the networks of the Society for Worldwide Interbank Financial Telecommunications (SWIFT), with the settlement and risk management functions being handled by correspondent bank relationships.

Returning to retail payments, one very popular payment method that we have not yet discussed is the credit card. This payment instrument dates back to 1915, with the establishment of "shoppers' plates" aimed a simplifying payment for affluent customers of retail establishments. It has evolved over the years into an enormously popular, globally acceptable payment instrument led by two major brand names: Visa and MasterCard. One of the principal reasons for the success of these two "card associations" is that they are owned and operated by a banks from all over the world. It is these local banks that manage the relationships with the cardholders; the card associations provide the global branding and also some of the common infrastructure that link the banks that operate the system.

Credit cards are used to make payments between customers and merchants. For the system to operate, the person making the payment must become a "card holder"—that is, he or she must approach a "card issuing" bank that must open an account on his or her behalf and issue the physical card that will enable him or her to make transactions. Without appropriate restrictions, the possession of a card confers unlimited spending power on its owner. In the majority of cases, though, the card issuing bank will assign a "credit limit" to the card holder based on an examination of his or her creditworthiness. In most developed countries, this process is quite routine; indeed, customers are often bombarded by advertising from different companies offering them credit cards. In other countries, credit cards are sometimes hard to get, and in some, tight restrictions are placed on their use.

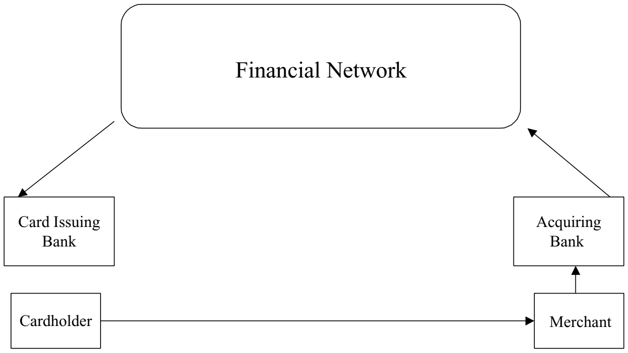


Figure 2: Information flow in a credit card transaction.

At the other side of each credit-card transaction is a "merchant." Achieving "credit card merchant" status involves opening an account with a bank that will "acquire" transactions on behalf of the business. Once the account is set up, the merchant then has the ability to charge arbitrary amounts to any credit card that has been issued anywhere in the world. Clearly, this represents a major opportunity for fraud in the short term, and acquiring banks will often subject a business to strict checks before permitting it to operate as a merchant, particularly if it intends to carry out business across the Internet. In the United States, these checks are not very stringent, but they are much more so in most European countries, and companies in some developing countries may have extreme difficulty in gaining credit-card merchant status.

Figure 2 shows the information flow when a credit card transaction is made. The cardholder presents the card details to the merchant. The merchant can authorize the transaction prior to actually making it. This is done through a connection either directly to the merchant's acquiring bank or to a technology provider acting on its behalf. The acquiring bank can authorize this transaction using a financial network that has access to the data of card-issuing banks worldwide. The transaction can have two steps—an authorization step (this is used frequently by hotels at the beginning of a guest's visit) and a later "capture" step where the previously authorized transaction is completed. Alternatively, an authorization-and-capture step can do everything in a single action.

It is, of course, not strictly necessary to make an online connection back to the source account for every credit card transaction. In situations where telecommunications facilities are not available, or where dial-up telephone connections are very expensive, the authorization step may just be a simple check of the credit card number against a periodically updated blacklist. Often merchants operate under quite complex policies to balance the risk of fraud against the cost of verifying the transaction. This may involve going through authorizations only where a transaction value exceeds a "floor limit" or carrying out an online authorization randomly for one in every 10 transactions.

One credit-card usage scenario that is interesting, because it serves as the background for Internet credit card transactions, is the so-called mail order/telephone order (MOTO) transaction. Under this scenario, merchants are allowed to accept orders by post or over the telephone with the customer simply quoting the credit card details verbally. Under this scenario—also called "card not present"—the merchant is unable to tell if the customer has the card in his or her possession. Neither can the merchant verify the signature. Some simple safeguards are put in place regarding the address to which the goods can be dispatched and in the event of the customer later disavowing the transaction, the merchant must bear the cost.

The costs involved in processing credit card transactions are considerable. Typically, these are recovered by a per-transaction levy on the merchant. The charges depend on the acquiring bank and also on the level of risk associated with the business. Typically, a merchant that has been trading profitably for years will be able to negotiate a better rate than a startup company. Any company that trades on the Internet is regarded as being "risky" and is typically subject to higher charges. Generally there is a fixed fee from around $0.10–$0.50 and a percentage of the transaction from around 1 to 5%. This effectively means that credit card transactions are not worthwhile for transactions less than $5 or $10.

The great strength of credit cards is their global acceptability. Since the processing of the transaction across the financial networks takes care of the currency conversion, merchants will receive funds in their local currency while the cardholder is levied in his. The global recognition of the two major brands and also other variants such as American Express reassures merchants that the payment will be honored. On the downside, fraud is quite easily perpetrated by rogue cardholders and rogue merchants,

Table 1 Consumer Preferences in Non-Cash Payment Methods by Country in 2000

COUNTRY	USE OF CHECKS	USE OF CREDIT TRANSFERS (GIROS)	PAYMENT CARDS	DIRECT DEBITS
U.S.	58%	4%	35%	2.2%
Netherlands	0.4%	39.7%	29.9%	29%
U.K.	26%	17.9%	36.6%	19.4%
Germany	3%	49.1%	9.6%	38.1%
Turkey (1997 figures)	6.9%	2.6%	83.9%	—
Namibia (1996 figures)	75%	14%	Not provided by local banks	9%
Angola (1996 figures)	75%	25%	Not provided by local banks	—

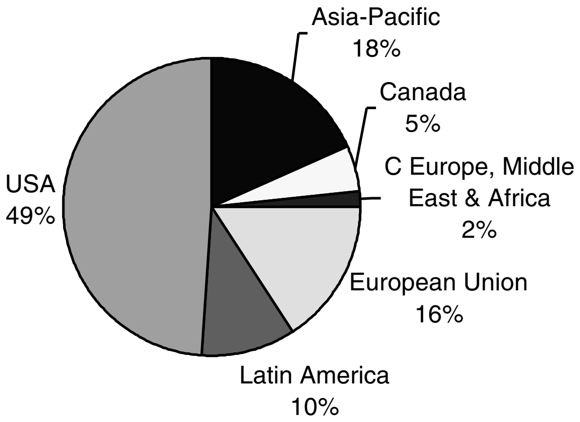


Figure 3: Geographic distribution of VISA cards issued around the world in December 1999.

particularly when the authorization process does not verify each transaction back to its source.

Geographical Variations

As Table 1 shows, the degree of adoption of different payment systems differs markedly between countries (BIS, 2001). The United States, for example, is a very heavy user of checks as a means of effecting payment. Payment cards come next in popularity, with quite infrequent usage of credit and debit transfers. It is almost the reverse in Germany, where payment card usage is extremely low compared with other developed countries and Giro credit transfers are used for more than half of the noncash transactions.

When developed countries are compared, many of the differences can be explained due to the historic evolution of payment systems over time. For example, the popularity of Giros in many European countries can be traced to the involvement of post offices in providing payment services over many years. In developing countries, the overall financial infrastructure tends to be poor. Since checks are perhaps the most basic payment instrument that a bank can offer, these tend to be available everywhere and attract widespread usage. In such countries payment cards have their use confined to particular industries (e.g., petrol stations) or they are not issued at all. Figure 3 shows how the 1 billion Visa cards that have been issued are spread throughout the globe. It can be seen that the area covered by Visa's Central Europe, Middle East and Africa division is responsible for just 2% of these. Often unusual local factors lead to a payments situation that is anomalous compared with similar countries. Turkey, for example, has embraced the use of payment cards to the almost total exclusion of other forms of payment.

ELECTRONIC PAYMENTS

In the development of e-commerce, the first transactions that took place were at a retail level, and this grew quite dramatically from just a trickle in 1995 to somewhere between $23 billion and $109 billion in 2000. Some product sectors that proved popular include books (e.g., amazon.com), apparel (Land's End, The Gap, Victoria's Secret), computer products (Dell, Gateway), and travel (Expedia, Priceline).

Credit Card Online Purchasing

In many ways, the most natural way to make a purchase over the Internet, in the absence of any alternatives, is to use a credit card. There was already a precedent set over a number of years by the catalog shopper. Business rules (the MOTO rules referred to earlier) had been developed to handle transactions where the card details were given to the merchant either on a printed order form, or over the telephone. For international shoppers, the currency problem was solved, and there were already large numbers of people worldwide who could make and accept payments without the need for any signup procedure.

When a credit card transaction is made online, the parties are exposed to the usual risks inherent in a conventional credit card transaction as well as some new risks that arise because the Internet is the medium of information transfer. These include the following:

Authenticity of the Seller: It is relatively easy for an individual to construct a credible online storefront that can appear to represent a large well-established organization. Even when the seller is authenticated, buyers who have given over their credit card details are at risk from rogue merchants who make repeat or completely bogus transactions against their accounts at a later time.

Authenticity of the Buyer: A credit card merchant has no dependable information on the identity of the persons making credit card transactions. They may have stolen the cards, or even generated valid credit card numbers at random, or they may be genuine cardholders who intend to later repudiate the transactions. Mechanisms such as cardholder address verification and the use of card verification values offer some protection against these attacks.

Security of the Information in Transit: Credit card details traveling over an unsecured Internet connection are vulnerable to capture from attackers intercepting the traffic.

Security of Information at the Merchant Site: Online merchants often capture the credit card numbers of their regular customers to ease repeat transactions. These databases are the target of attackers and there are reported cases of thieves compromising this information and using it to extort payments from merchants.

Cardholder Privacy: Archived details of online transactions form a link between a person's identity (as represented by his or her credit card) and the goods he or she purchased. Merchants can use data mining techniques to build profiles of their customers, which may represent an unwanted invasion of privacy.

Funds Availability: As with any credit card transaction, it is possible that a rogue cardholder may ultimately default on payment to the card issuer. In conventional commerce, the merchant would be indemnified against this risk, but for online transactions, the credit card companies may take a different view.

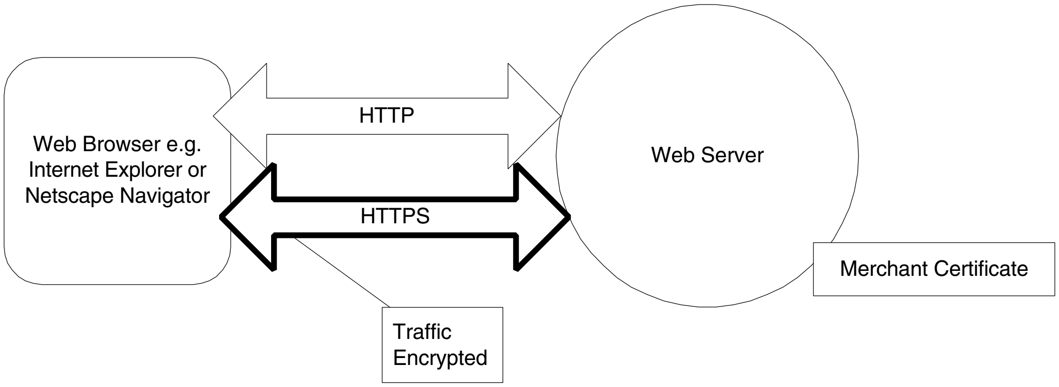


Figure 4: The secure socket layer protecting a credit card payment.

The Secure Socket Layer

The earliest Web purchases were made either by insecurely transferring the credit card details in a Web dialogue or by resorting to a separate e-mail exchange to complete the payment. The credit card companies were not happy about this method of conveying the details, and the advice they issued to consumers and merchants was not to use credit cards on the Internet until such time as new technologies had been developed to allow this to happen securely. The market largely ignored this advice and a stopgap solution was arrived at around 1995, when Netscape incorporated support into their Internet browser software for a technology standard called the secure socket layer (SSL) (Dierks, 1999).

The secure socket layer allows systems to authenticate themselves using digital signatures and X.509 Certificates. Digital signatures are produced by electronically digesting the document to be signed and producing a small unique piece of data that represents an electronic fingerprint of the document's contents. This fingerprint is then encrypted using a secret number called the private key. The encrypted fingerprint is the "digital signature" and the only person that can sign a document is the holder of the private key.

When any other party wants to verify that the signature is correct, they decrypt the document using a nonsecret number referred to as the public key. If the result matches the fingerprint of the document they accept the digital signature as genuine. In order to use this technique as the substitute for a real signature, the last thing we need is a way to associate a person's identity with a particular private key.

When people travel from one country to another, they assert their identities by producing passports. These documents provide a link between their appearances (and their handwritten signatures if necessary) and their identities (name and date of birth). They are accepted at border posts because passports are issued by national governments trusted by those officials.

The electronic counterpart of this is called an X.509 certificate. It is an electronic document that provides a link between a public key and an identity (person or company name) and is signed by an entity called a certification authority (CA) that is widely trusted.

Once we have a single CA or an international network of cooperating CAs that is widely trusted, it is possible for people to send signed digital documents such as electronic checks to each other. By including the X.509 certificate with the document, they allow the recipient to verify the signature on the document.

A merchant wishing to use SSL to protect credit card transactions must apply to a recognized X.509 certification authority to be issued a certificate. All Internet browser software comes preconfigured to trust the 20 or so most common certification authorities operating worldwide. As Figure 4 shows, a user browsing the merchant's site will interact normally until it comes to the point where the credit card details are to be transferred across the link. At this point, the user's browser will be directed to a Web page that starts with HTTPS rather than the usual HTTP. This is a signal to the browser to start a special security dialogue with the browser in which two things happen. First, the merchant proves that it represents the business to which the X.509 certificate is issued, and second, the merchant and the user agree on a session encryption key that is used to protect the credit card details and any other financially sensitive information from being intercepted by attackers as it travels across the Internet.

Thus the cardholder is afforded some protection that the merchant to whom he or she is giving the card details exists as a bona fide business, at least as of the time when the certificate was issued. Both cardholder and merchant are also protected from eavesdroppers capturing the credit card details from an insecure Internet link. There is no protection for the merchant from the card being used by someone other than the holder, and if the cardholder later denies making the transaction, there is no means of proving otherwise. The cardholder has no protection against a merchant who may retain the card details and subsequently change multiple transactions against the account. If the merchant site stores the card details online, it makes itself vulnerable to attackers breaking into the site to gain access to them.

In order to streamline the process of making credit card transactions and also to allow each individual transaction to be authorized, merchants generally equip themselves with online connections to their acquiring banks or to entities operating on their behalf. Figure 5 shows this process in action.

This process has been taken further by companies such as iTransact that operate links to the financial network

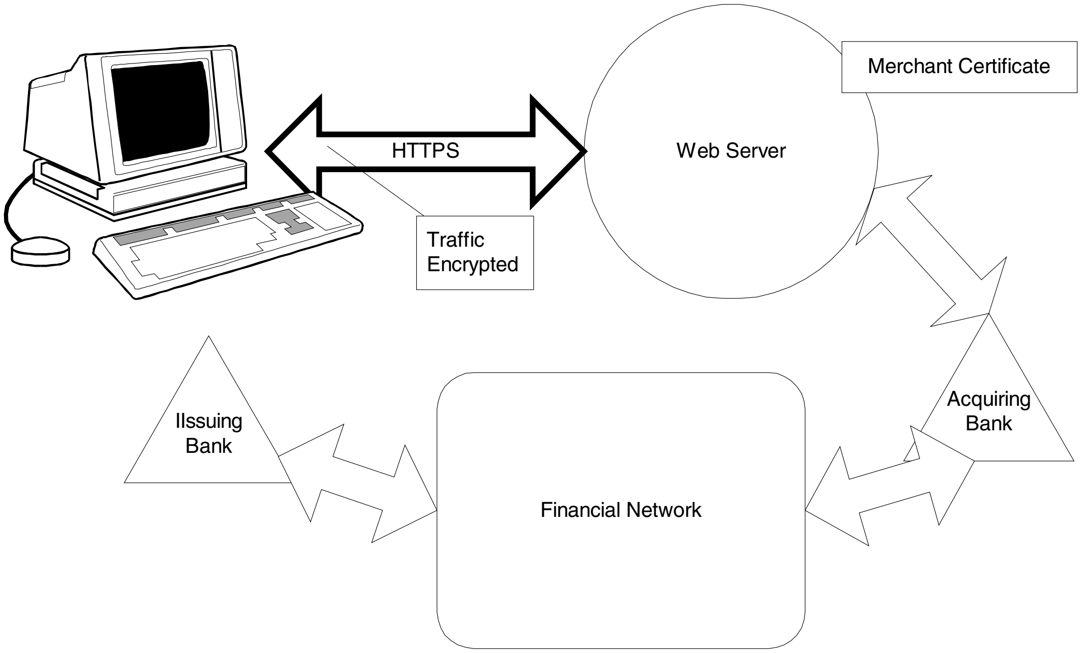


Figure 5: Online authorization of a credit card transaction.

on behalf of many hundreds of online merchants. Using their services, the B2C merchant can interact with iTransact's Web site during a purchase to authorize and complete a transaction in real time. The only requirements are that the merchant hold a merchant bank account in the United States and that the transactions be denominated in U.S. dollars. Every other component of the system, including the merchant Web site, can be located elsewhere.

Secure Electronic Transactions (SET)

Although the use of SSL, with or without online authorization, is by far the most common means of making credit card transactions, more advanced technology is available in the form of a security standard called secure electronic transactions (SET) (Mastercard, 1997). This was developed principally by the two major credit card companies, Visa and Mastercard, in 1996, but also has the support of many major technology providers and other card brands such as American Express. It is a comprehensive solution to all of the practical risks that are encountered in any credit card transaction.

Special wallet software is used by the cardholder that is partially or totally integrated into the Web browser's software. The wallet software is loaded with the card details and also with a certificate that is issued to the cardholder by the issuing bank. As shown in Figure 6, when a credit card transaction is to be made, the wallet software composes an encrypted payment request that is sent via a SET module running on the merchant's Web site and from there to an SET payment gateway run either by each acquiring bank or by the credit card company itself. The SET standard underwent a 1-year public review period and is thought to be highly secure and efficient at guarding against all anticipated risks due to stolen cards, rogue merchants, or rogue card holders.

The main problem with SET lies in its complexity. Three independent pieces of software need to be in place and working well with each other before a single transaction can be carried out, and certificates must be issued to each of the three parties to allow them to securely identify each other. Banks began to pilot SET at the beginning of 1997, but this was done mostly on a regional basis (which does not fit well with the global way in which the Internet operates), and these pilots achieved limited success in terms of persuading large populations of users and merchants to change over to the new system. As of mid-2002, SET has still achieved little market penetration and its proponents are beginning to experiment with so-called "light" versions of the standard that involve less complexity.

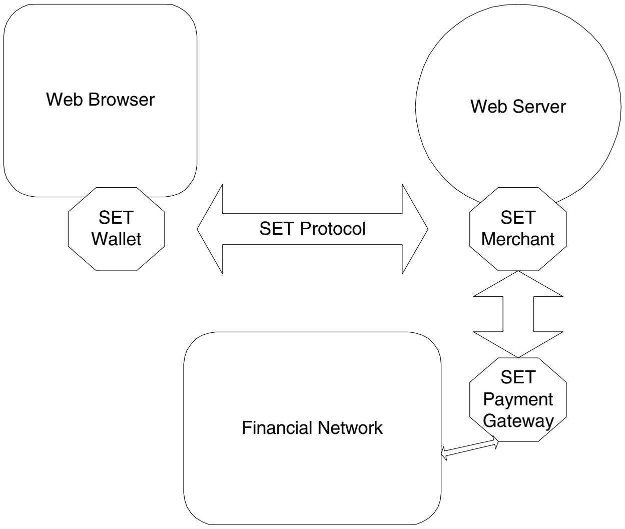


Figure 6: The secure electronic transactions protocol.

Account-Based Systems

Most of the technology involved in a credit card transaction arises because the merchant and buyer have accounts with different banks (possibly in different countries) and each transaction involves both a check for funds availability and ultimately a transfer between them. All of this can be greatly simplified if all users hold accounts with one entity. At the point when a transaction is to be made, this one entity is contacted and requested to transfer the funds from the buyer's account to the seller's account.

There are numerous system of this type available on the Internet, most of which are operated by companies that are not banks or financial institutions. Two that are worthy of mention due to their substantial customer bases are Yahoo PayDirect and PayPal. These systems all have links to conventional payment systems, e.g., bank accounts or credit cards to inject money into or withdraw it from the system, but conceptually any method can be used. Indeed some systems, e.g., PocketPass, allow the account to be primed with cash by buying a prepaid card in a store.

The difficulty these systems have in succeeding is that they essentially create a private currency that is only acceptable to people who have registered to use their services. Since, initially, very few merchants accept the currency, this makes it less attractive for buyers to sign up to use the system. This chicken-and-egg problem has caused many companies to fail in providing payment services.

Cash-Like Systems

Since cash is used for approximately 80% of the transactions that people make, one would expect that there would be a great demand for this service for electronic commerce transactions. Market reaction to some of the earlier cash-based systems (most notably eCash from Digicash BV) was less enthusiastic than expected. This company launched and deployed a software-based system that allowed individuals to make arbitrary fully anonymous transfers of value between each other in a range of currencies. E-cash was rolled out in many countries around the world in conjunction with local partners (e.g., in the United States with Wells Fargo bank and in Germany with Deutsche Bank), but in most cases was not a big success. The company has now refocused on a portfolio of payment solutions including a person-to-person transfer method.

Many other systems that claim to be cash-like are available in the market, but for the most part, these fall into the category of account-based systems, where the payment is simply a transfer between identified accounts on the provider's system.

Smart-Card-Based Systems

In conventional bank-mediated transactions, the trend for retail point-of-sale systems is away from paper-based instruments such as checks and toward electronic payment effected with a card. Most of the cards in use today are based on magnetic strip technology with some rudimentary account identifying information recorded (insecurely) on a magnetic strip on the back of the card. The banking industry is in the process of transitioning to the next generation of payment cards based on so-called smart-card or chip-card technology. Here, the plastic card has a chip mounted on its surface. When inserted into a card reader, this chip powers up and is able to have electronic dialogues with the card reader device.

The advantage is principally that the card can carry much larger amounts of information in a form that cannot be copied. The chip on the card encrypts data before sending them to the card reader, making it very difficult to break the security. Secret quantities such as cryptographic keys never leave the card. On the down side, the cards are more expensive to produce and are vulnerable to attacks from card reader hardware that has been subverted.

One application of the chip-card technology is in the realization of an electronic purse. Value is loaded into the smart card for later spending. There are two main efforts on-going in this area, the first by Mondex International and the second by a consortium led by Visa called the common electronic purse specification (CEPS). Of the two, the Mondex effort is more mature and has been in common use since 1992. The Mondex system offers a means of transferring value from one card to another. A person can transfer value from his card to that of his friend by simply inserting both cards into a hand-held value-transfer terminal. Similarly, bricks-and-mortar merchants can use a point-of-sale terminal containing a merchant card into which the buyer inserts a Mondex card to allow a transfer to take place. The Mondex card is in use in over 50 countries around the world, including several in sub-Saharan Africa. Pilot experiments have been conducted to use this system to make purchases across the Internet, but no large-scale scheme has yet been attempted.

The Visa-sponsored CEPS system is at a much earlier stage, with the CEPSCO consortium having been formed in late 1999, and early implementations are just beginning to appear. One of the difficulties of using smart-card-based payment methods for e-commerce is that each user's terminal must be equipped with a smart-card reader. Although many thought that this hardware would become part of a standard specification PC, this has not yet happened. It would seem that the prospect of smart-card-based payment methods becoming important in E-commerce is extremely unlikely for at least 2 years and possible indefinitely.

Where smart cards may play a role is in the emerging area of mobile commerce (m-commerce). Since all GSM digital phones contain smart cards (referred to as subscriber identity modules or SIMs) and there are expected to be a billion mobile phone subscribers in the world by 2002, this represents a large user base. As yet, though, it is far too early to tell what form the mobile Internet will take and whether the presence of a SIM will be influential in determining how consumers make payments in this environment.

Bank-Mediated Payments

Electronic Checks

The Financial Services Technology Consortium (FSTC) is an organization made up of the main American banks and banking technology providers formed in 1993 to enhance the competitiveness of the U.S. financial services industry

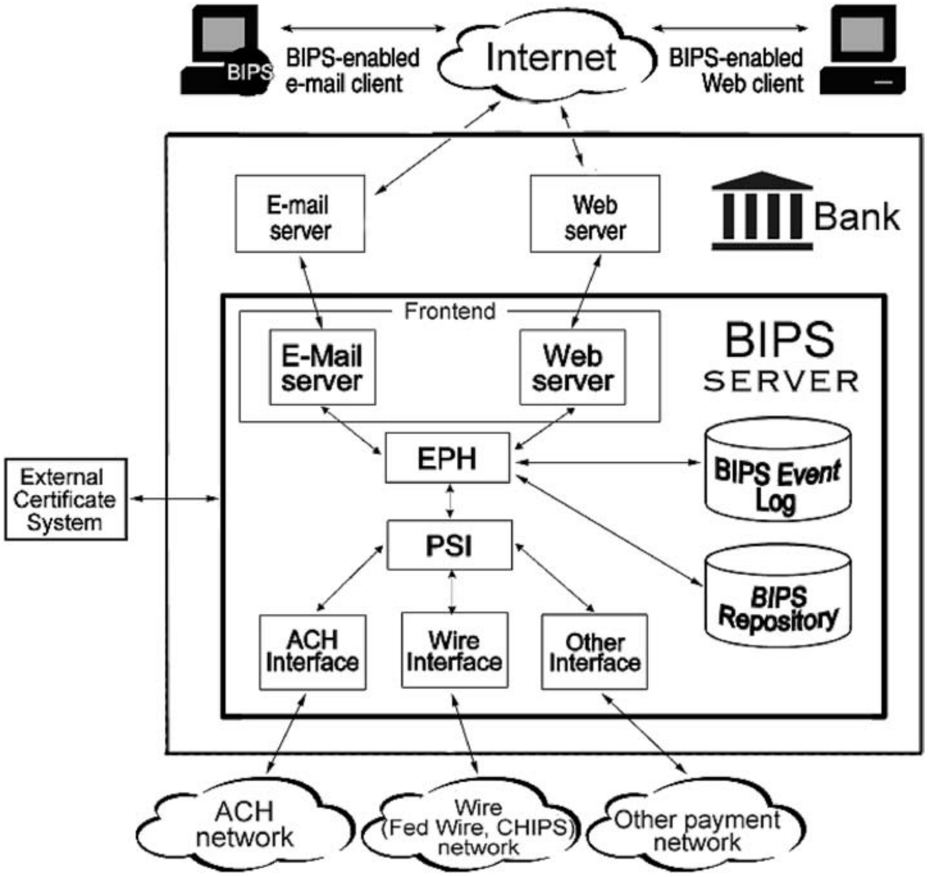


Figure 7: The bank Internet payment system (BIPS) architecture.

through the use of technology. For some years, it has been working on a specification for electronic-check-based payment (FSTC, 2001).

An electronic check is a document containing fields identical to those that might occur on a paper check, with appropriate digital signatures being added when the check is first issued by the payer and also when it is endorsed by the payee. A limited pilot using these checks was run in mid-1998, and subsequently, a new syntax for expressing the check was defined, called the financial services markup language (FSML).

Electronic checks expressed in FSML may be exchanged by trading partners in future B2B exchanges. Before these can be processed, though, the banks involved must have appropriate technological infrastructure to process them and use the information contained within to effect the required interaccount and interbank transfers. The FSTC has also laid out an architecture for upgrading a bank's existing technology to add the capability to handle electronic checks issued and transferred between organizations on the Internet.

FSTC Bank Internet Payments System (BIPS)

In 1996, the U.S. Financial Services Technology Consortium (FSTC) initiated a project to come up with a very general way to allow companies easier access to payment services. Their approach involved making as few modifications to existing U.S. banking systems as possible. Figure 7 shows how the Bank Internet Payment System (BIPS) acts as an Internet "front-end" to the existing ACH, Wires transfer, and other bank networks. Messages such as "payment requests" can be initiated by either e-mail or Web-based software. A public key infrastructure (PKI) is assumed to exist, and this component could be provided by the Identrus consortium, who are building an international PKI for banking applications.

BIPS was demonstrated in a number of projects involving the Glenview and Mellon banks in the United States in August 1998, but there has been no public progress beyond that.

Mobile Payments

In many European countries, most notably in Scandinavia, the percentage of the population who use mobile phones has exceeded 70%, and most other parts of the developed world are following the same trend, with industry estimates suggesting that there were more than 1 billion mobile phone users worldwide in April 2002.

Since each mobile phone can be viewed as a computer with network access, it makes an ideal payment instrument. In addition, GSM mobile phones are equipped with subscriber identity modules (SIM) that can be used to identify phone users and encrypt traffic between them and the network. A final plus point is that phone subscribers usually have billing arrangements with their network operators to which online purchases can be charged.

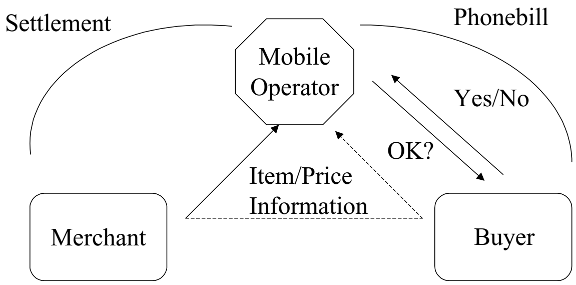


Figure 8: Making payments with a mobile phone.

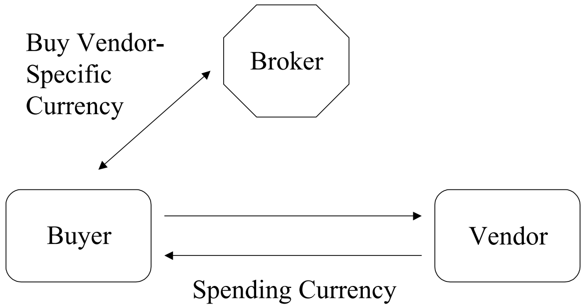


Figure 9: Typical micropayment system.

Many different payment systems that involve phone users making purchases using variations on the simple steps shown in Figure 8 are in development.

The buyer and merchant agree on the goods to be purchased and a reference to this information is conveyed to the mobile operator, possible via the buyer. The information movement can take the form of a phone call to a specially designated number or a short message or text message or may be part of a dialogue using the wireless application protocol (WAP).

In order to make the payment, the buyer is usually prompted by the mobile operator to indicate acceptance of the offer. Once again, this may be done using text messages or more elaborate dialogues, but it leverages the fact that the operator has already positively authenticated the buyer and can charge the amount of the purchase to the buyer's phone bill. The mobile operator will then perform periodic settlements with the merchant, usually deducting a percentage of the purchase price as a commission on the sale.

Micropayments

For some classes of goods and services, for example, access to information, the payment of a single once-off amount is inappropriate. In conventional commerce, this is typically handled by a subscription arrangement. Individuals subscribe to a magazine to be able to gain access to the articles published over a period of a year. While it might be interesting to sell access to a single article in a magazine, the value of such goods would be so low that it would be uneconomical to charge and collect such an amount.

Micropayments are a technology that makes such payments economical. They focus on allowing the payment of repeated very small amounts (e.g., 1 cent or less) to a single merchant. In a typical payment (called a macropayment), a buyer communicates with a seller, often using cryptographic mechanisms to secure the link or to perform authentication before a payment is made. The seller then typically communicates online to a bank or payment provider to check that funds are available and to complete the transaction. For micropayments to be economical, we must eliminate two aspects of this. First, strong cryptography cannot be used between buyer and seller, as to do so for such small transactions would mean that the seller's computer systems would be very limited in the dollars per hour of goods that they could sell. Second, we need to eliminate the online check to a third party when a transaction takes place, as there are very finite limits on the number of connections per hour that a third party's computer systems can handle.

Most micropayment systems involve a third party called a broker who issues a type of electronic currency bound to a particular vendor. Generally, this currency will be formed using lightweight cryptographic techniques and will use hash algorithms rather than more heavyweight encryption algorithms such as DES or AES.

As an example, at the beginning of the week, a buyer may make a credit card payment to purchase $20 worth of currency specific to a newspaper publisher. Throughout the week, he will then spend this currency as he browses individual articles in the newspaper. This process is shown in Figure 9.

Any unspent amounts can be redeemed at the broker. Then newspaper publisher will periodically redeem the spent micropayments at the broker and have its account credited.

In order to qualify as a micropayment, it must be possible for the vendor to check the validity of the currency without performing any very complex cryptography and without needing to contact the broker. Generally, it is impossible to do this without leaving some opportunities for fraud to occur. Micropayment systems typically allow a limited amount of fraud, but since the amounts being dealt with are small, the financial exposure is limited and worth the risk.

One well-developed micropayment was the Millicent system developed by Digital Equipment Corporation in 1995. A broker system issued a vendor-specific currency called "scrip." For each transaction, the buyer handed over the scrip and was given the appropriate change as a new "scrip" of lower value. This system underwent commercial trials in Japan in 1999, but has not progressed to wider usage. Other notable micropayments include the Payword system developed at MIT and IBM's μ3-KP.

Micropayments have yet to undergo any very large-scale deployments. One possible reason for this was the development of online advertising on the Internet. The ability for content providers to be paid tiny amounts based on page impressions for online advertising provided a solution to the problem that micropayments targeted and robbed them of their natural market. The slump that has taken place in online advertising, coupled with the fact that there are many other applications for micropayements, may see this technology deployed in the future.

CONCLUSION

In conventional commerce, many different ways of making payments have evolved, but most of these are based on the principle that the bulk of the payments occur locally. For example, a conventional check payment clears quickly when made to a merchant who shares the same bank. It is a little slower when there are two different banks operating in the same country and clears very slowly indeed when made between countries. The conventional payment method that works best globally is the credit card, and this had carried over well to the Internet, where very often the buyer and seller are very far apart. The bulk of Internet payments today are made using credit cards across a secure socket layer, which, although it protects against some of the risks involved, is still far from ideal. Many alternative technologies exist that can improve upon this, but they have proved difficult to deploy.

Many diverse payment methods exist that have characteristics similar to cash and check payment, as well as novel methods such as micropayments and payments initiated from mobile phones. While these methods often have many desirable properties that are improvements on conventional payment methods, they have yet to see broad deployment. Probably the major reason for this is the lack of locality on the Internet. It is difficult to assemble an initial population of buyers and sellers that can be persuaded to sign up for a new system and reach the critical mass required. Attempts by banks to introduce SET in particular areas failed due to this chicken-and-egg problem. Buyers will not sign up until there are a wide range of merchants using the system and merchants will not sign up unless there is a customer base that they can target.

It is likely that we will see deployments of many of the systems discussed above in the future, but it may take many years to establish a standard and it may require the backing of companies who have sufficient global presence on the Internet to reach the critical mass required within a very short period of time.

GLOSSARY

Credit cards Used by charging against the customer credit and are by far the most popular method used in electronic payment systems.

Micropayment systems Include stored financial value for online payments and are used for small payments such as pennies and fractions of pennies.

Secure socket layers Allow systems to authenticate themselves using digital signatures and X.509 Certificates.

Smart cards Include stored financial value and other important personal and financial information used for online payments.

CROSS REFERENCES

See *Business-to-Business (B2B) Electronic Commerce; Consumer-Oriented Electronic Commerce; Electronic Data Interchange (EDI); Electronic Funds Transfer.*

REFERENCES

Bank for International Settlements (2001). *Clearing and settlement arrangements for retail payments in selected countries*. Retrieved June 1, 2001, from http://www.bis.org/publ

Dierks, T., & Allen, C. (1999). *The TLS (SSL) protocol Version 1.0, IETF request for comments RFC-2246*. Retrieved January 1, 1999, from http://www.ietf.org/rfc

Financial Services Technology Consortium (2001). *FSTC electronic check project details*. Retrieved December 1, 2001, from http://www.echeck.org

Mastercard Visa Corporations (1997). *Secure electronic transaction (SET) specification: Book 1. Business description Version 1.0*. Retrieved May 1, 1997, from http://www.setco.org

O'Mahony, D., Peirce, M., & - Tewari, H. (2001). *Electronic payment systems for e-commerce* (2nd ed.). Norwood, MA: Artech House.

Electronic Procurement

Robert H. Goffman, *Concordia University*

HISTORY OF E-PROCUREMENT

The concept of e-procurement has technically been around since the 1970s but in different formats and data transmission vehicles. The first electronic transactions took place through the use of punch cards and data phones that functioned as a primitive modem. A data phone transmits data one-way to prepunched cards identifying certain descriptions and quantities of the order. This system worked well with standard quantities preset by the manufacturer or supplier and usually required the merchandise to be stock items. The typical ordering consisted of maintenance type items that were regularly stocked in the supplier warehouse.

The next form of electronic procurement did not emerge until the early 1980s with the development of the facsimile machine (fax). The fax machine also provided one-way communication but offered several significant advantages over the data phone. First, it required printed material, which allowed for more detailed information than the data phone, including drawings, specifications, prices, and so on. Second, the fax did not require a computer technician to input the order and therefore was accessible to everyone. Both of these technologies had drawbacks including the following: (a) historical data was not kept electronically, (b) there was no communication assurance, (c) there was no security in place, and (d) they required manual lookups. The only way to verify receipt of the order was to follow-up with a phone call.

With the growing use of personal computers (PC) came the use of diskettes to send and receive data. A purchasing agent placed the appropriate purchasing information on a diskette and mailed it to the supplier. The supplier "read" the information, filled the order, and sent the diskette back to the buyer with the status of each item ordered. The use of diskettes provided a mechanism for maintaining historical information, increasing accuracy, and providing purchasing confirmation. Using the PC in conjunction with a fax machine further helped expedite orders. For example, a buyer could fax an order to the supplier then send a diskette with back-up detail and for tracking purposes.

This use of the PC for procurement became the foundation for electronic data interchange (EDI). As computers became more prevalent and powerful in the mid-1980s, EDI began to flourish. An EDI system is used to transfer purchase orders (POs) via electronic format from the buyer to the seller. It eliminates the necessity for double entry (data input from both the buyer and the seller) and allows additional information to be exchanged electronically, for example, forecasts of requirements, supplier shipping notification, overdue purchase orders, requests for quotations, material certifications, statistical quality analysis, and order changes and cancellations (Neef, 2001, p. 55). As EDI became an increasingly common form of procurement, technical standards were created to establish consistency among suppliers and buyers.

The Internet explosion of the mid-1990s created new strategic and technological visions in many areas. Use of the Internet and Intranet for e-procurement became a viable solution for corporations because it functioned as a powerful communication vehicle for transmitting data between buyers and sellers. The feasibility of using a full e-procurement solution increased in the public and private sectors as Internet browsers, such as Netscape and Internet Explorer, became user-friendly and the use of computers became routine.

From the late 1990s through today, the extranet has become the new strategic selling point for companies. It allows corporations to leverage their capabilities, extending volume discounts to their partners, receive additional revenue for use of internal suppliers, and offer price reductions.

INTRODUCTION TO MODERN-DAY ELECTRONIC PROCUREMENT

Today's e-procurement may be the next technological revolution for corporations, governments, and consumers worldwide. It became a technological priority for most top executives during the e-business boom era (1996–1999) and continues to be on many corporate technological agendas, even today. E-procurement has proven itself at some of the prominent corporations in the United States and also efficiently supplies the U.S. Navy. EG&G Logistics partnered with SupplyWorks to resolve the navy's problem of not having all supplies at the base's supply store when the ships came to port. With this partner solution, navy personnel now have the ability to order online while at sea through SupplyWorks' Web site using an Internet browser. The orders would be completed and delivered to the base store by EG&G ready for the personnel when they arrive.

Another prime example of an early adopter of e-procurement is General Motors (GM). GM began seeking an e-procurement solution in late 1999 and implemented its first version the following year. With a $63 billion procurement expense, GM expects substantial cost savings through the use of e-procurement. Savings would be captured through reduced prices, better analysis and information, better supplier control, and so on. It was also a logical transition as GM was already familiar with and using EDI. In early 2000, the automotive industry made a significant announcement regarding procurement. All three major U.S. manufacturers were about to embark on the creation of a single Internet-based procurement network. The purpose was to lead, control, and benefit from the significant technological advancements that were taking place.

A large part of e-procurement savings comes not from the reduced transaction costs and prices, but through strategic information and negotiations. Database management is a large part of e-procurement. "A fundamental characteristic of sourcing decisions is their uncertainty. A buyer never has complete information about all aspects of supplier performance and [its] future development" (Essig & Ulli, 2001). The goal for most e-procurement solutions is to provide more information in order to gain substantial benefit in varying ways, including cost savings, negotiating power, more accurate forecasting and analysis, reduction of uncertainty, and mitigation of risk. These are clearly reasons why GM wanted to enter into e-procurement sooner, rather than later.

With savings already being recognized by organizations such as GM and the U.S. Navy, e-procurement came to the forefront of media hype referencing how the e-business revolution would change the corporate world, and the way business is conducted fueled the fire. Industry analysts suggested the world was about to become "virtual" overnight, thus eliminating the need to visit stores or physically handle merchandise before purchasing. All shopping would occur electronically over the Internet, and companies that spent years solidifying market share would remain viable only by moving away from brick-and-mortar to selling over the Internet. Such prognostications certainly increased sales for those whose technologies facilitated Internet access. Nothing happens overnight, however, and efficient and effective use of these hyped products remained unproven. Moreover, several of the "revolutionary" products and technologies that were promoted as providing e-commerce turned out to be nonexistent or overrated. Ironically, those consumers who purchased goods over the Internet encountered a variety of problems including limited security and customer service. Through hindsight, we know that the majority of e-commerce companies still in business have survived by skillfully managing their cash flow, are well funded, and have a real product that buyers want.

There are a variety of ways in which people currently use the Internet for commerce. This chapter focuses on business-to-business (B2B) e-commerce as it specifically relates to e-procurement. This segment of Internet commerce is growing with total transaction values estimated to grow to $2.2 trillion by 2003 and $7.4 trillion by 2004 (Rybeck, 2001). Although the bursting of the technology bubble has caused companies to pull the purse strings on many new technology initiatives, most corporations continue to view e-procurement as a promising option and a part of their future.

What Is Electronic Procurement

E-procurement is a subcategory of e-commerce and supply chain management. At a basic level, it involves purchasing merchandise or services through a Web site. At a more global level, e-procurement provides an efficient, paperless solution for purchasing goods and services by allowing the end user to order an item from a computer or other internet connected device, such as a cell phone, kiosk, television top box (i.e., Microsoft's X-Box), and so on, and using an electronic payment system while storing necessary information for future analysis.

Definition

The Center for Innovation, Business, and Manufacturing (CIBM; n.d.) in Australia nicely summarizes e-procurement: "E-Procurement is the use of the Internet to connect buyers with suppliers to facilitate the purchasing of goods and services focusing on business-to-business (B2B) and business-to-government transactions (B2G)."

An alternative, more detailed definition provided by the Aberdeen Group (2002), a Boston-based consulting firm, defines e-procurement as follows: "Internet-based procurement (e-procurement) creates private, Web-based procurement markets that automate communications, transactions, and collaboration between supply chain partners." Aberdeen Group research indicates that e-procurement offers the most direct and dramatic means for organizations to reduce costs, improve productivity, and increase profits (Aberdeen Group, 2002).

Although other ways of implementing e-procurement have been available for years, use of the Internet as the electronic transmission vehicle currently is the key element that characterizes e-procurement. Prior to the current technological advances, special software and connectivity was necessary for e-procurement through the use of EDI and proprietary packages. The Internet allows for the masses to use an inexpensive, easy-to-use, procurement method through a Web browser. Today, the newer e-procurement solution methods use the Internet, eliminating the complexities of EDI and proprietary setups and software. "However, in late 2001, most observers would concede that more talk than transaction has flowed through the Internet enabled 'supply chain of the future'" (Aberdeen Group, 2001).

SUBDIVIDING E-PROCUREMENT

Some analysts further subdivide e-procurement into indirect and direct. These terms refer to the type of transaction occurring and differentiate between items bought to run a business (indirect) and those purchased to make a product (direct).

Indirect Procurement

Indirect procurement involves purchasing any item required for daily activity. Common types of indirect transactions include office supplies and equipment, travel, replacement parts, and so on. In a typical corporate environment, indirect procurement accounts for more than 60% of all purchases, typically transacted by the end user (Neef, 2001).

Indirect procurement may be further subdivided into ORM and MRO. ORM is the products and services used to facilitate daily business routines such as office supplies, travel, furniture, computers, and printers. These types of purchases usually involve high-volume, low-cost items that are purchased on an as-needed basis. MRO includes those purchases necessary to keep production running, such as replacement parts and servicing of manufacturing equipment. Ordering these products is more complex than ORM because of varying volumes, the need for regularly scheduled services, and the critical nature of the purchase.

Direct Procurement

Direct procurement is any purchase a company makes that "directly" relates to the production of the product it sells. For example, for a company manufacturing computers, direct procurement would include purchasing hard drives, memory boards, circuits, CD-ROM, and so on. Direct procurement is often a part of supply chain management (SCM). SCM is any solution that automates procurement and activities related to buying the raw materials, parts, and assemblies necessary for manufacturing finished products.

Direct procurement can be more complex than indirect procurement for MRO purchases because the company often buys materials in large volumes from a limited number of vendors through a limited number of transactions. In addition, materials may be specialized and "made to order," creating even more complexity. Direct procurement is "mission critical," and an increasing number of firms will be developing e-procurement strategies specific to such purchases. In particular, experts anticipate an increasing focus on direct procurement strategies among manufacturing firms where procurement can account for up to 60% of their total procurement expenditure (Neef, 2001).

Implementation Strategy

Because MRO solutions are much more complex than ORM implementations, many companies pursue the ORM strategy first and follow-up with the MRO and direct purchase initiatives as a second phase.

The advantages of this type of phased implementation are twofold. First, immediate financial savings can occur by implementing an e-procurement plan related to purchasing "noncontract" ORM products such as office supplies. By purchasing contracts and establishing prenegotiated discounts with preferred vendors, a company can stop the average employee from shopping anywhere and save significant money. Of course, policies preventing maverick buying must be put in place and enforced to gain the full benefit of supplier negotiated contracts. Second, the non–mission critical nature of ORM products allows an organization to "practice" or "test run" strategies without jeopardizing the heart of its business. This reduced risk provides the company with time to learn and modify its plans and activities. Furthermore, a technology solution for purchasing office supplies and equipment will most likely be easier, quicker, and more efficient than a single comprehensive e-procurement solution for all areas.

This phased implementation approach may have varying scenarios, depending on the type of organization. Risk-averse organizations may want to take small steps and prove functionality and technical capabilities before a large undertaking. In this case, one may implement a view-only catalog (to be addressed later in this chapter) for a small group of users. Once this functionality is proven, another piece may be added to the solution, and so forth until a complete e-procurement solution is established.

Although this methodology requires more time and money to implement, many companies would have greatly benefited from this approach during the e-business boom and crash. This risk-averse approach may have easily saved some companies millions of dollars by realizing that the proposed solution was not viable early in the project. With less dollars being spent on technology and a more cautious business environment, most companies are now using this phased approach to technology projects.

Combining Direct and Indirect Procurement Into a Single Entity

Most organizations have focused on e-procurement for ORM. Realizing cost savings, time savings, and increased process efficiencies, e-procurement solutions are now being considered and implemented for MRO. Industry experts are currently forecasting a consolidation among direct and indirect procurement over a long period of time. Although this merger may not happen for a decade or more, great benefits will be realized once it occurs. Procurement organizations will be streamlined, proficient, and better leveraged. The procurement area will no longer need to divide its business and resources in two. Efforts will be focused on strategic sourcing and procurement negotiations. E-procurement solutions will continue to drive or contribute to reduced product prices, increased productivity, and overall cost savings.

The consolidation of ORM and MRO procurement would be extremely difficult without an electronic solution in place. The business process, negotiations, and people are significantly different between the two entities. An e-procurement solution allows the purchasing manager to spend his or her time where it is most needed, not deciphering a person's handwriting for a particular order.

E-PROCUREMENT ARCHITECTURE

Because there are multiple entities involved in implementing a full e-procurement solution successfully, the architecture varies from company to company and depends on what part of the process an entity is responsible for (i.e., buyer, supplier). Some general basic scenarios are identified in this section, as well as technology considerations. There are also some diagrams to help understand architecture at a conceptual level.

Buy Side Scenario

The architecture for a buy side model begins with the electronic catalog. The e-catalog is a conglomerate of all products and services from all approved suppliers and is located on a company's intranet. This is where the employees will shop and order products. The e-procurement software allows for electronic order approval, creates order forms for each supplier, and routes the order to the appropriate individual supplier. See Figure 1.

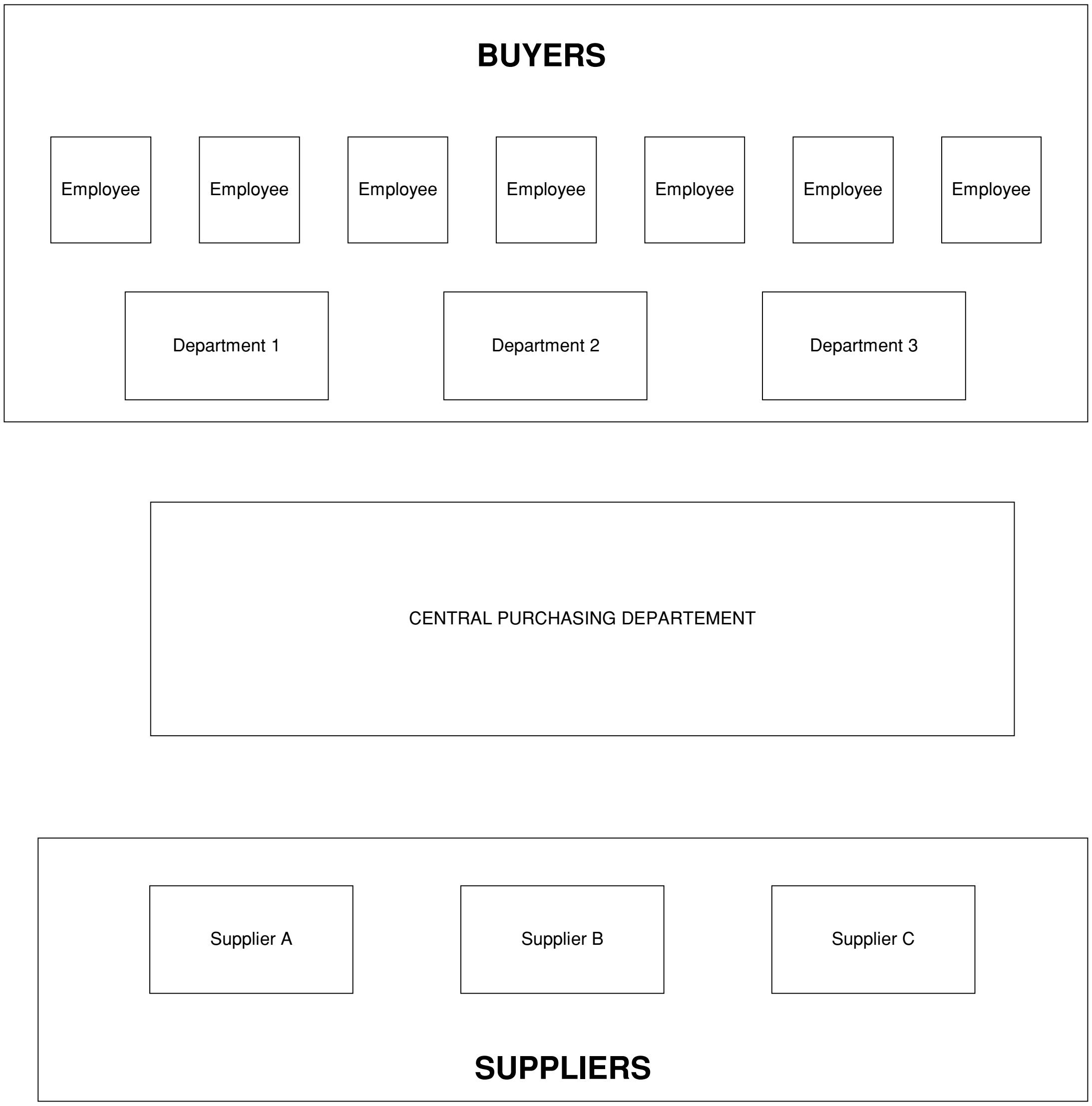


Figure 1: Procurement flow with central purchasing department.

The heart of e-procurement in the buy side model is the e-catalog. Without an easy to use, efficient, quality e-catalog, customers will not return to the intranet site. Customers will make their purchases through an alternative method, if possible.

The power behind an e-procurement solution is the software. Selecting the right vendor and software is also critical to success. The software should integrate into the current infrastructure and support the requirements gathered in the analysis. Although the buy side implementation may be complicated, the ultimate savings are worth the effort.

Sell Side Scenario

Most consumers are familiar with the sell side model. Shopping on the Internet at Web sites such as amazon.com and toysrus.com are examples of the sell side model. The suppliers offer items for sale on the Web site and receive direct orders from the consumers.

At a more complex level, suppliers integrating with large corporations may interface order processing with the client's back-office applications. Integrating supply chain management systems with e-procurement allows suppliers to replenish inventory automatically and reduce possible backlogs. More recently, suppliers have also added functionality to update the customer with real-time order status information via e-mail. See Figure 2.

Just like the buy side model, the heart of the sell side model is the e-catalog. The catalog gives the supplier visibility and drives customer business. Technically, the sell side model is usually easier to implement, but it is important to make sure that the same standards are in use at the supplier and buyer and that quality and service levels are monitored closely.

Electronic Catalogs

One of the most difficult pieces of implementing an e-procurement solution is the catalog. Establishing an effective e-catalog requires significant planning as well as search capabilities for item comparisons. In addition, maintenance of the system can become a major concern. Who will maintain the catalog? Where will the catalog be hosted? How often will the catalog be updated? What will be the process for maintaining the catalog? How are special requests handled? Most implementations focus on transaction efficiency and neglect the content management process until late in the project.

The complex and time-consuming task of developing e-catalogs has brought a new enterprise to market: content aggregators. The content aggregator works with suppliers to gather information and deliver it in a standard, searchable, user-friendly format (Intel, 2000). Some specialists in content management, aggregation, and delivery are Aspect, Requisite, and Harbinger. In addition to corporate e-procurement, many content aggregators also have e-marketplaces for their customers.

As mentioned in the buy and sell side models, the heart of e-procurement is the e-catalog. A successful solution will entice purchasers to return again and again, be easy to use, efficient, and effective. One bad experience will make it difficult to regain a customer.

Technology Considerations

Standard e-procurement architecture does not exist. Each organization most likely has a technical infrastructure already in place. E-procurement solutions will need to be integrated into current infrastructures or implemented within a larger technical solution such as an enterprise resource planning (ERP) solution.

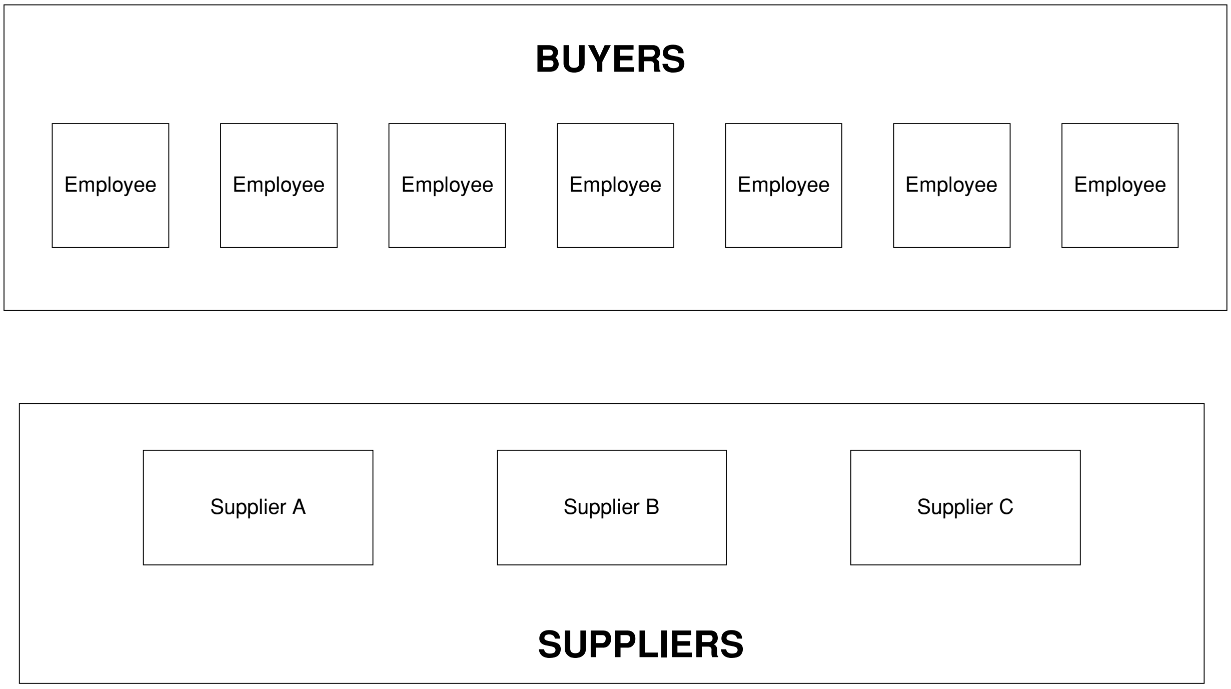


Figure 2: Procurement flow with empowered employees.

E-procurement architectures will vary from organization to organization and will certainly be different between suppliers and buyers. Because initiatives and business processes are different among suppliers and buyers, the solutions will need to adapt to current strategies.

There are some basic strategies about architecture. One will need to consider if there will be multiple vendors or a single vendor, if the solution will be integrated with other systems or set as a stand alone application, how the transmissions will take place, and whether all purchasing will take place through the electronic solution. Also, it is necessary to determine if Internet portals will be considered as a part of the solution.

When evaluating the architecture, there are several important questions to ask as well. What is the current organizational culture? Is a centralized or decentralized purchasing process in place? Should the e-procurement solution be a centralized or decentralized purchasing solution? The answers to these questions are critical to designing the e-procurement technical architecture and future business process.

To be successful, it is critical to evaluate a company's business process and what changes will take place. Make sure to have input from all stakeholders, especially the end user. This will not only help ensure success, but make the implementation much smoother and quicker.

Payment Systems

From a technical perspective, one of the most difficult pieces to implement in an e-procurement solution is the automated payment system via electronic transmission. A lack of formal standards accounts for the majority of concerns in this area. Over the past two decades, EDI has been the method for electronic transmission with standards in place. Unfortunately, only firms with deep pockets can take advantage of EDI because of the high expense and complexity entailed. In 1996, however, a newly developed standard called extensible markup language (XML) promised an affordable solution for secure transactional business data between firms via the Internet. A variety of vendors such as SWIFT and ACORD work to provide XML formats as an alternative system.

XML continues as the replacement method for EDI. In 1998, the Data Interchange Standards Association (DISA) conceded that XML as a Web-based technology would likely replace the traditional American National Standard Institute (ANSI) X12 EDI as the business-to-business standard for business data exchange. XML standards today are individually structured by industry, however, and therefore are not consistent across sectors. For example, the financial services industry uses the XML standards set by the SWIFT Cooperative, whereas the insurance industry uses XML formats set by ACORD. Even across software vendors XML standards are different. ARIBA uses cXML (Commerce XML), whereas Commerce One uses xCBL (XML Common Business Library). Although the basics of XML are intact, the standards still vary. For a financial services company to transact with an automotive company, middleware needs to translate the messages. The ultimate goal, however, is for a single standard of XML to evolve that provides standardized and predictable dictionaries of XML terms and repositories that store and manage the product descriptions (Neef, 2001).

PROCUREMENT STRUCTURE IN SUPPLY CHAIN MANAGEMENT

E-commerce is an extremely broad area that increases in scope on a daily basis. E-procurement is one area within the realm of e-commerce under SCM, a concept that encompasses all activities surrounding buyer–supplier relationships from forecasting and production planning through procurement to customer service and order management. The Supply Chain Management Group identifies the end points of SCM as the customer's customer and the supplier's supplier. Procurement is one small, but costly, piece of the supply chain. There are two market approaches to procurement: vertical and horizontal.

Vertical Markets

Vertical markets focus on one particular industry, such as financial, telecommunications, or transportation, and may cater to a core group of clientele. Many suppliers, especially in the manufacturing sector, are structured as vertical markets for niche items. Some of the leading hubs are chemdex.com, esteel.com, and neopharma.com. In addition to providing e-procurement services, vertical hubs often provide current news and other content of interest on their Web sites. Many buyers use vertical market suppliers out of necessity either because they require niche products or because of the specialized service. A vertical market supplier understands a particular industry in depth and works closely with its buyers to improve products and processes from both the buyer and supplier perspective. It is often a very close partnership. The hub provides similar services and benefits of an in-house e-procurement solution, but on a much smaller and much more limited scale. Simply stated, a vertical market hub gives the buyers direct access to member suppliers thus reducing product costs through the elimination of the middleman.

Horizontal Markets

Horizontal markets focus on broad categories that cross multiple industries such as software, electricity, or utilities. Although the majority of e-markets are structured vertically, horizontal markets are beginning to expand. Some examples of the newer horizontal markets can be found at redtagbiz.com, bidtag.com, and adauction.com. An excellent example of a horizontal supplier is an office supply store, such as Office Depot or Office Max. All companies require pens, pencils, paper, and so on. Office supply stores do not focus on any particular industry but distribute to all types of companies. Other, less tangible types of horizontal suppliers are gas and electric companies. These companies provide a basic need to everyone, regardless of the business involved.

Like the buyers and suppliers of the vertical market hubs, horizontal hubs allow direct access to suppliers therefore reducing costs. The supplier advantage of becoming a member of a hub is the increased visibility and potential for greater sales. Companies may need to use both vertical and horizontal markets. Just as indirect and direct procurement are slowly moving toward a single entity, vertical and horizontal markets will probably merge and become transparent as more companies implement e-procurement. This merger may be a decade away but is both feasible and likely.

THE PROCUREMENT PROCESS TODAY

For the majority of companies, the procurement process today is still manual and paper-driven. Supervisor and manager approvals are often required as well as a cumbersome procedure for ordering everyday items. Today there are two approaches to procurement, decentralized and centralized.

Centralized Procurement

Prior to the mid-1980s, most companies had a centralized procurement schema. All purchases had to go through a single purchasing department (Figure 3). For larger, multilocation companies, there may have been a purchasing office for each location or region, or have a single procurement department at the headquarters where all items were approved and ordered. Although many government agencies have a centralized procurement area they are beginning to realize the benefits of e-procurement and are considering electronic solutions. (See G2B: The Government and E-procurement later in the chapter.)

Although the centralized process may appear cumbersome, there are key advantages:

- Control—the organization retains purchasing knowledge and can better manage expenses.
- Contractual vendors—all vendors are contractual. Factors such as prices, quality, specifications, and service level agreements (SLA) are prenegotiated.
- Standardization—order forms, policies, and procedures are uniform across the company.
- Efficiency—orders are bought from appropriate vendors with invoicing and payment usually being centralized in the purchasing department.

The primary disadvantage to centralized procurement is time. Because of the multiple layers of approval, purchasing from an employee's perspective is often long and cumbersome. A purchasing department may delay ordering until there is enough demand or need for a particular product. Although this may help reduce cost through bulk purchasing, overall cost is higher when taking into account lost productivity, probable maverick buying by the end user, and additional effort tracking items purchased, on hold, and rejected. These disadvantages are easily eliminated through e-procurement and decentralization.

Decentralized Procurement

In the late 1980s, the paradigm began to shift toward decentralization. Companies wanted to streamline the process and possibly realize some time and cost savings through decentralized procurement (Figure 4). Although still a paper-driven process, a person or manager was now empowered to order low-cost items without approval. There are many strategies and levels of decentralization. Some companies allow invoicing and payment to become a department responsibility, whereas others leave this to the procurement department. The procurement area's responsibility ranges from duplicating orders on a shadow basis (the order is entered twice but only sent once) to complete autonomy and focus on strategic sourcing and vendor management. This responsibility depends on the need and desire for control at a corporate level.

THE E-PROCUREMENT DECISION

E-procurement solutions exist for both centralized and decentralized organizations. If the organization is centralized, all orders placed by the end user would be sent

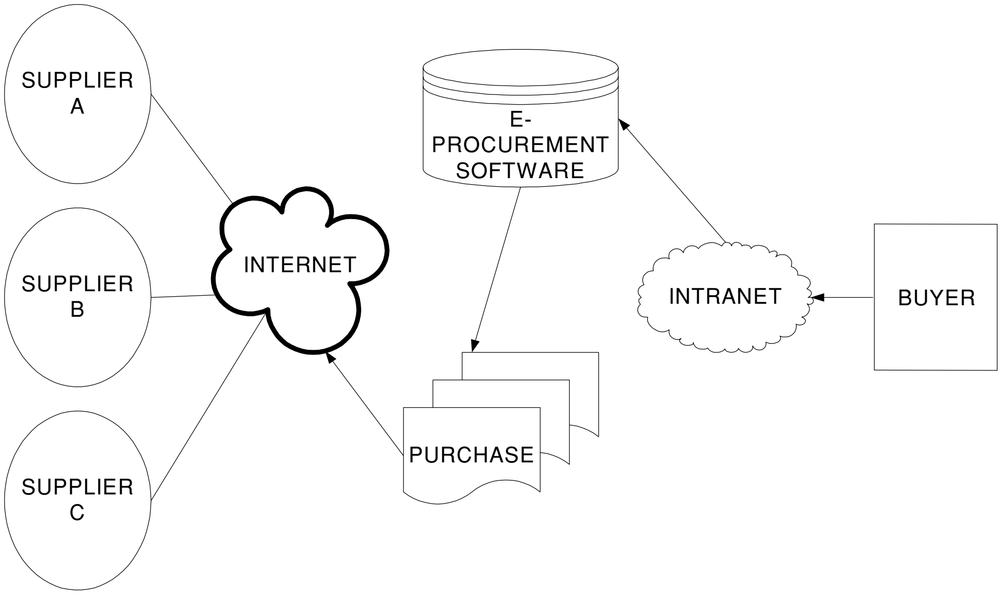


Figure 3: Generic centralized e-procurement architecture.

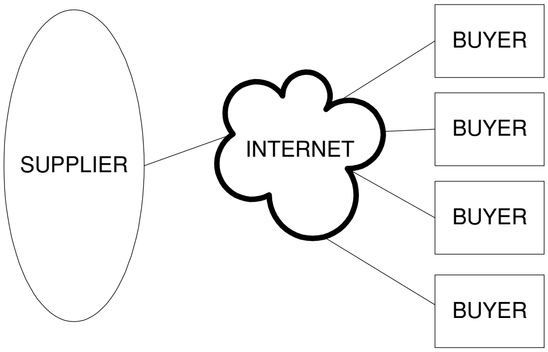


Figure 4: Generic decentralized procurement architecture (employee empowerment).

to the centralized purchasing location for review and approval. E-procurement streamlines the process, reduces cycle time, and lowers cost through use of negotiated contracts, reduced maverick buying, and process efficiencies gained. In a decentralized organization e-procurement efficiencies are gained through e-catalogs, reduced order errors, reduced cost through use of contractual suppliers, and improved productivity. Both structures can benefit by e-procurement solutions. There are several key considerations to review and analyze before deciding to design an e-procurement system.

Evaluation of E-procurement

Advantages of E-procurement

As almost every article, book, and publication on the topic notes, e-procurement benefits the bottom line in many ways. The main advantage of e-procurement is reduced costs through various avenues, including the following:

- Improved internal efficiency—productivity will increase through the use of e-catalogs and automated approval, eliminating paper catalogs and laborious, manual, paper-driven ordering and approval processes.
- Minimized maverick buying—employees will reduce or eliminate buying supplies from noncontractual vendors.
- Increased purchasing specialists bargaining power—the primary role of the purchasing specialist will no longer be order taking and "translating" but deal making. The specialist will negotiate volume discounts based on the entire company's budget. The specialist will focus on limiting the number of suppliers and ensure SLAs.
- Cut supplier costs—suppliers gain efficiencies and lower costs by reducing overhead through automating tasks previously done manually, such as order scheduling, advance shipment notice, and order acknowledgment.
- Reduce order error rate—with the inception of e-catalogs, new product and pricing information can be updated immediately, reducing order error rate. Suppliers will focus on value differentiation instead of order problems.
- Increased product visibility—the e-catalog increases vendor and product visibility by use of search engines, Web sites, electronic markets, and Intranet catalogs.
- Electronic catalogs—e-catalogs increase user productivity through reduced search time and better presentation and information. The e-catalog also substantially reduces supplier print and distribution costs.

Challenges of E-procurement

Although e-procurement has the potential to save an organization millions of dollars in a single year, the solution does not come without great challenges, including strategic, cultural, and technological obstacles. Some of the main challenges companies have encountered are the following:

- Rethinking old assumptions—Web sites must be interactive and enticing, not just content providers. Create a site that will entice users to return to it.
- Business process reengineering—current manual processes and procedures will no longer work. New business processes need to be created and implemented including the following:

  Streamlined approval process,
  Reduction in suppliers,
  Monitoring of suppliers for quality and service level,
  Negotiate discounts with suppliers,
  Use of e-catalogs and new software applications,
  Order taking,
  Inventory tracking and reporting,
  Shipping, and
  Billing and payment.

- Reinvest in technology infrastructure—servers and telecommunications will now need to handle substantial databases and Web site usage efficiently. Systems should scale in processing capability, memory, and input–output requests.

It is clear that the largest challenge when implementing an e-procurement solution is the business process. The process, policies, and procedures are re-created and affect all personnel. This is why focus on corporate culture and business process is arguably more important than the technology itself. Training is a great expense in time and money. With a new procurement process and system in place, a company will need to train all users (probably the entire company) and will need to create a rollout plan. For most companies, this means maintaining duplicate procurement processes for some period of time. Some people will continue to use the old method until trained while trained personnel will use the new system. There are of course, ways to minimize the time period and duplication, but more likely than not, some period will be required.

One other area of concern and possible challenge when making the e-procurement decision is the financial consideration. Is there a budget for the solution, does the financial analysis validate the executive decision?

Financial Considerations of E-procurement

There are many ways of performing financial analysis when determining if e-procurement is monetarily the correct solution. Some methods to consider are net present value (NPV), return on investment (ROI), and payback period (how long will it take to recoup the cost).

Since the e-business bubble began to burst in 1999, most companies are more cautious with their technology investment dollars. E-procurement solutions are expensive. Depending on the size of company, cost may be in the millions or tens of millions of dollars. This should not discourage one from undertaking an e-procurement initiative. The financial analysis will most likely demonstrate that implementing an e-procurement solution is well worth the initial investment and maintenance cost. Remember, in addition to the visible dollars saved, there are many additional savings via "soft dollars" such as increased productivity and efficiencies.

The financial consideration and budget are important considerations in addition to all other analyses previously discussed. Combining all strategic, technology, and financial considerations together, the executive team is now ready to make a decision.

Corporate Strategic Procurement Considerations

There are many reasons to pursue an e-procurement solution within a company, including the ability to find new sources, lower transaction costs, shorten cycle times, reduce prices, and provide more value added activities. Of course, successful e-procurement strategies must also take into account the need for security, training, time, money, and standardizations. Depending on the size of the organization, global, national, and local strategies may need to be considered.

Global Procurement Strategy

Procurement strategies, policies, and procedures should begin at the enterprise level of an organization. For large, international companies, this means evaluating at a global level. Executives must analyze current procurement practices and evaluate the complexity, benefits, cultural challenges, and financial impacts of how an e-procurement solution will affect the overall strategic direction. Will an e-procurement initiative affect any ongoing or planned activities? Does it make sense to implement e-procurement throughout the entire enterprise? Will all departments accept a new procurement solution readily, or will cultural impact need to be addressed? These are some questions to ask when evaluating globally.

Enterprise strategies, analysis, and decisions, are often performed by senior executives. These executives view the entity as a whole and are not involved in detailed departmental activities on a regular basis. Therefore, participants from a national, regional, and local level must be included in the final decision and plan.

Intel provides a good example of a global procurement strategy and mentions some of the difficulties with a worldwide implementation. Two challenges Intel will need to address are standards for connectivity and bandwidth. With the hundreds, if not thousands, of other customers its suppliers have, placing standards around connectivity for e-procurement will prove challenging. Although bandwidth in the United States should not be a concern, there are countries where bandwidth and infrastructure will not be able to handle the procurement solution.

Regional Procurement Strategy

Similar to the view of global procurement strategy, the personnel responsible for national or regional strategies will also evaluate at a conglomerate view. Current procurement practices at a global level are limited; however, many companies do have regional procurement centers. Evaluation and analysis of contemporary and future procurement practices will be more detailed regionally with current buyer–supplier relationships and policies and procedures for purchasing in place.

Current culture and business processes must be considered when looking to the future. Again, will employees accept e-procurement as a solution? Will the corporate culture need to change? How will the business process change? What risks will the business units endure? Although these questions are similar to those for the global evaluation, they need to be addressed and answered with a more hands-on procurement perspective.

Local Procurement Strategy

Finally, there is the local procurement strategy, often considered the office or department. Senior management often gives the directive for local strategy. Although the local strategy is predefined by regional and national levels, it is still important to include local stakeholders in the decision process. These frontline personnel will often have substantial qualitative input. They are also the final stakeholders of the end result.

Strategic Technology Considerations

Global Technology Initiatives

It is important to evaluate all projects and technology plans in conjunction with the e-procurement solution when determining an overall strategic direction. Performing a thorough analysis up front will alleviate some major headaches later. For example, if a firm is currently implementing an ERP system solution, decision makers should ensure that e-procurement integrates into the current technology plan and system. It is also important to plan for upgrades, compatibility, scalability, and possibly advanced future technologies, such as integration with personal digital assistants (PDAs).

Executives evaluating global technology initiatives often view the enterprise from a limited perspective. Senior management makes global decisions based on research, subordinate input, media hype, budget, and coordination with other technology projects. To gain a better user and benefits perspective, it is important to have regional and local personnel involved in the decision-making process and plan.

Regional Technology Initiatives

Technology projects are often initiated and implemented at the regional and local levels of an entity. Globally, there are few technology projects that benefit an entire corporation. One primary example of a global technology project is ERP, which usually spans the enterprise. Although e-procurement can certainly benefit the entire enterprise, it is often initiated within an office or region. Again, current technology and planned projects must be evaluated when considering e-procurement. One must be certain that the decision is in line with the strategic technology plan and that it will meet expectations. Once a pilot of the e-procurement solution is proven, executives may want to expand it to the entire company.

Local Technology Initiatives

Local considerations are similar to regional technology initiatives; the main difference is functionality and perspective. At the local level (offices or departments) the view is much more from an end-user perspective. This is mainly where the front-end analysis is performed. What is the current procurement process? How can the procurement process be improved? What functionality is required for an electronic solution? These are the types of questions to ask at the local level. Technology personnel are also excellent resources to identify local challenges, type of employee reception an electronic solution is likely to encounter, and type of training that may be required. It is important to include stakeholders at all levels when considering a new technology initiative.

THE DECISION—DOES E-PROCUREMENT FIT?

When evaluating the final decision, there are two critical paths of questions to ask when considering e-procurement: organizational questions and technical questions. The decision approach will most likely be made by a top-down approach. In other words, the executive management team will view the initiative at the enterprise level. The executives then work with other management at the regional and local levels to ensure that e-procurement is the right initiative. Are there higher priorities? Is there confidence in the technology? Is the organization culturally ready for a complete business process change? These are just some of the initial questions to answer when making the final decision.

Cost-cutting measures are certainly a high priority on the short list, and e-procurement is one solution to reducing tangible and intangible costs over a short period of time. The challenges are substantial in many ways, but there is usually no reward without taking a risk. If e-procurement fits into an organization, it should be embraced and installed. If your organization is not ready for e-procurement, do not endure the technological difficulties and high risk.

For Purdue University, e-procurement was a fit. After 30 years of paper, Purdue spent $4 million implementing an e-procurement solution that went live in late 1999–early 2000. Purdue's immediate expectation of savings is 7% through more favorable supplier agreements and a reduction in staff. There are many other savings and benefits that will be realized over a longer period of time, such as better information, better control of expenses, negotiating leverage through automation, and more (Ariba Solutions, 2001).

BUILD OR BUY?

The build versus buy decision should be a "no-brainer." There are many quality and proven vendors with expertise in e-procurement. Because of the extreme complexity and scale of an effective solution, using current vendors in the marketplace is highly recommended. Building a solution in-house is unlikely to keep up with the ever-changing technology environment and standards set by suppliers, buyers, and vendors. Documentation of in-house systems is usually nonexistent. Support is usually poor, but that can also be the case by choosing the "wrong" vendor.

The more difficult question to address is whether to hire a services firm for installation and integration or to use in-house personnel. Supporting commentary by both Purdue and Intel say it is difficult to predict ROI and difficult to compare the cost of outsourcing or installing in-house. The major factors involved here are expertise, personnel, and time.

Buyer Perspective

The authors of *E-Purchasing Plus* identify 10 points to consider when starting an e-procurement solution:

1. Build your team.
2. Scan the environment.
3. Obtain management support.
4. Tie e-purchasing to your broader e-business strategy.
5. Perform supply chain audit and review.
6. Promote speed and convenience.
7. Manage and improve supplier relations and performance.
8. Change, change, change.
9. Maintain a strategic focus.
10. Create a sales interface. (Giunipero & Sawchuk, 2000, p. 113)

Additional steps necessary in conjunction with these 10 items in preparing for this large, complex project are the following:

1. Establish a management environment: leadership, vision, commitment, change process.
2. Identify the customer.
3. Understand the procurement customer and needs—internal (chief executive officer, chief financial officer, engineer, maintenance) and external (suppliers).
4. Determine procurement strategies.

There are many advantages to promote when trying to gain acceptance of an e-procurement solution. Some of these include savings through reduced product prices and lowered procurement costs. These savings come through several channels, including vendor negotiations, reduced maverick buying, straight-through processing, and elimination of paperwork and extensive labor and approval processes. Another advantage is increased user satisfaction. Employees will be empowered to purchase products necessary for daily routines and enjoy other benefits such as less hassle and quicker service, response, and delivery. Multiple suppliers and catalogs may be transparent to the end user, increasing efficiency and time savings. All of these advantages lead to more productivity while reducing costs.

In addition to internal advantages, suppliers increase their benefits and should pass some of their cost savings to customers. Sales may increase, visibility with end users will improve, supplier cost will be reduced via electronic processing, errors will diminish, and an electronic payment process will reduce transaction and payment delays. These are just some of the advantages to begin promoting when considering e-procurement. Another step in gaining acceptance is to evaluate the requirements of the Web buyer and the purchasing manager.

The requirements between an individual consumer and a company using e-procurement vary somewhat. Web buyers want speed, integration, information, ease of navigation and navigation options, and a service contact person. If the system is too cumbersome or the Web site does not work efficiently, consumers will want to purchase their product elsewhere. Purchasing managers, on the other hand, want to empower the end user to purchase standard and low-cost items (indirect purchasing), training them on free navigation, electronic catalog, privacy and security, purchase limits and restricted access, search capabilities, access for authorized suppliers, and so on. The front-end or user-interface is geared toward the Web buyer and the back-end or inner workings of the system is geared toward the purchasing managers requirements. Both of these needs must be met for the solution to prove successful.

In the past, small firms have been barred from e-procurement because of cost. In today's environment, however, smaller firms may have two options: using a marketplace or possibly using a financial services firm such as their bank. With regard to the former option, more and more vertical markets are popping up on the Internet. Ventro Corp. has created marketplaces for the life sciences industry. Some of their ventures are Promedix, specializing in medical supplies; Broadlane, specializing in health care commodities; and Industria, supplying pipes, valves, and fittings. The next option is leveraging a bank or other lender. Many banks, such as Wells Fargo, Wachovia, Bank of America, MasterCard, and American Express have created alliances to help smaller firms gain cost savings through their corporate contracts. For example, Wells Fargo has a large contract with Boise Cascade as its supplier. For a small membership fee, Wells Fargo's business clients will be able to leverage the prices within the Boise Cascade/Wells Fargo Agreement and purchase via Wells Fargo. American Express gives their corporate clients discounts at various places through their contracts, such as Federal Express and Office Max. Use the card and the discount is automatically applied.

So, what's the advantage for the financial service firm? Revenue. The bank can charge a membership fee; however, the real revenue would come through loan financing. For example, if a Wachovia business customer wanted to purchase 100 computers through Wachovia's supplier, financing can be arranged immediately through the bank. In this scenario, Wachovia provides the loan and also collects the membership fee. It also creates a stronger barrier for a customer to finance elsewhere.

Supplier Perspective

As previously stated, many suppliers implement SCM systems into their organization. It makes sense to use this current system for e-procurement because it establishes direct access to suppliers and reduces costs. In addition, the SCM may have add-on modules for e-procurement or may already be integrated into the system. An alternative route is to purchase a solution that will integrate into the current infrastructure. As mentioned earlier, building a solution will take more time and money than the effort is truly worth.

E-procurement Software Specialists and ERP Vendors

The number of e-procurement specialists on the market today is still limited. Several large name companies (e.g., Ariba, Itwo, Commerce One, and ProcureNet) and a few smaller corporations (e.g., Concur Technologies, Clarus, and Trilogy) dominate the specialist market. Today many of the ERP software companies are beginning to incorporate e-procurement modules into their solutions as well. Some big names in this arena include SAP, Oracle, and Peoplesoft. Initially, e-procurement software focused on the purchase of indirect goods, but within the last year, advances have been made to include direct purchases as well. Also, some of the larger e-procurement players such as ARIBA have created portal-based communities to which companies can subscribe and participate.

Ariba, Commerce One, and Purchase Pro are known to have built-in sourcing tools that support request for quote (RFQ) and the negotiation process. Other products, Maximo and Get-It, although strong in partner information and integration capabilities, may be limited in the sourcing area. Clarus has strong RFQ and auction tools. Often a company will receive a recommendation in choosing a product through either a current relationship already in place with a supplier, consulting firm, or logistics firm.

G2B: GOVERNMENT AND E-PROCUREMENT

A predominant area for e-procurement initiatives is in government organizations, the military, universities and hospitals. It is believed that e-procurement will not only save these organizations millions of dollars a year, but also will increase efficiencies, bargaining power, and service. Many of the purchasing procedures remain manual and require approval by managers, directors, or vice presidents, depending on the request. E-procurement solutions may automate the entire process, therefore eliminating many, if not all, of the approval steps and substantially reduce process time.

One of the largest areas of government spending is the military. E-procurement initiatives are becoming common place in the defense sector, although this may be due to the suppliers rather than to initiatives among government purchasing departments. Major defense suppliers such as Lockheed Martin and General Electric (GE) are streamlining their own procurement processes and moving to e-procurement systems with defense system buyers following suit. In addition to saving costs and increasing efficiencies, e-procurement allows supplier bidding to be more competitive, thus driving the material purchase prices down.

A further result of government's move toward e-procurement has been establishment of Web portals for purchasing and contracts Some example Web sites are http://www.fedmarket.com, http://www.gsaadvantage.gov, and http://www.nigp.org. An additional benefit is increased government accessibility and availability of information for contract announcements, awards, requests for proposals, procurement announcements, and bid notices.

Although areas of the public sector are extremely interested in e-procurement solutions, especially from a cost-savings perspective, many are slow to make a decision. Often there are monetary and political considerations. Although this solution has the potential to save organizations millions of dollars per year, there is also a large upfront cost to implementing the solution. Many organizations do not have the monetary capacity or budget to consider a full e-procurement solution today.

Because e-procurement has been proven in the private sector, many public, government, and nonprofit entities are beginning to consider e-procurement as a viable cost-reduction solution and planning future budgets to incorporate e-procurement into their organization.

THE FUTURE

E-procurement has made substantial technological advances over the past few years and is expected to continue improving efficiencies and productivity in the future. Consolidation is taking place by the solution vendors and will continue over time. Ultimately, the vendors that will remain will need to provide fully scalable integrated solutions—not only for electronic procurement, but most likely for the entire supply chain management structure.

E-procurement will become more virtual in the sense that employees may be able to purchase from any Internet-connected device and have supplies delivered to any location overnight. Users will request additional functionality for e-procurement such as enhanced interfaces with internal systems, automation for time and expense, support for electronic invoicing and payment processing, and improved reporting functionality.

The e-procurement solution will integrate seamlessly into an organization's e-commerce strategy and help enhance productivity as well as increase corporate profitability. "E-Procurement offers the greatest opportunity to improve processes, increase productivity, and reduce costs across the supply chain" (Aberdeen Group, 2001).

BEST PRACTICES

Over the past two years, the Aberdeen Group affirms that e-procurement is delivering "significant and verifiable benefits." Organizations are required to rethink fundamentally their sourcing, supplier management, and procurement strategy and processes when considering e-procurement. Proven implementations have overcome these challenges and provided the beginning of e-procurement best practices. Planning a strategy for e-procurement consists of new business processes and strong management directive and includes the following:

- Setting supplier selection criteria—reduce the number of suppliers through a methodical selection criteria approach. Using a minimum number of suppliers enhances negotiating power and increases savings.
- Planning for scalability—as catalogs and electronic usage become more robust, the infrastructure must be ready to handle large volumes of data and transactions.
- Reviewing suppliers' content update strategies—accurate information is a necessity for corporate buyers. Content management is critical to success.
- Choosing compatible solutions—Use your current infrastructure. If an ERP solution is already in place, work with your vendor for compatibility.
- Increasing information technology bandwidth—higher intranet and Internet usage may put a strain on the system. Ensure high network availability by expanding bandwidth.
- Building redundancy and load balancing—redundancy and load balancing traffic will also help ensure network availability. Slow, unresponsive systems lose customers.
- Evaluating all options—perform a thorough analysis of all business and technology requirements. Evaluating suppliers, vendors, and other stakeholders will help ensure a successful implementation and solution.

When executives contemplate if or how to implement an e-procurement solution, it is extremely important to include the overall corporate strategy, culture, and architecture. The majority of system implementation failures are due to a lack of incorporation and "buy-in" from all stakeholders. Several initiatives that address this nontechnical aspect of e-procurement implementation are the following:

- Improve efficiency and reduce labor costs by eliminating the manual, paper-based processes and providing enterprise-wide, self-service procurement;
- Enforce on-contract buying; eliminate maverick buying;
- Gather accurate and meaningful data on total spending, both by supplier and type of purchase (decision support);
- Using supplier performance, select preferred suppliers for strategic sourcing;
- Move as many transactions as possible to front-line employees without undermining business rules; and
- "Smooth out" the supply chain: integrate process and systems, internally and with suppliers (Neef, 2001).

A large part of success depends on a quality communication plan with the end user. More often than not, a company implements a new system and then simply expects everyone to use it. Regardless of the technology or the successful implementation of the system, plans like this almost always result in failure because they do not meet the users' needs and expectations. The end users in these situations typically do not use the new system or change their old habits unless some form of coercion takes place, such as eliminating the original system. With regular communication, everyone remains informed about the progress and expectations, thus fostering acceptance by all parties involved with or affected by the project.

Another key to success is engaging the appropriate mission-critical resources, such as procurement managers. Obstacles to this process depend on the people involved. Many managers are wary of e-procurement systems, fearing they will lose their job or power. To mediate this type of situation, depending on the size and complexity of the organization, executives need to reassure the procurement specialists of their job security and emphasize the critical role they will play in the purchasing process. Procurement specialist will not be eliminated, but reinvented; they will no longer manage paper catalogs and mounds of paper purchasing requests and invoices. They may manage an electronic catalog and content, create and oversee purchasing policy development in regard to the new e-procurement solution, and ensure compliance with policies implemented. Their positions will be more critical, not less.

The procurement specialist is the logical choice for the "super user" of the system and will be the expert for decision support, catalog management, vendor management, service level negotiations, and so on. Ultimately, this person should be cutting costs and maintaining metrics to assure executive management that the e-procurement solution is delivering the expected success. As described earlier, efficient and effective communication between all the people involved plays a critical role in the successful implementation of a new system. It is imperative for executives to remember this human side of the business equation.

CONCLUSION

The e-business boom brought e-procurement to the forefront of technology. E-procurement continues to be on everyone's mind, especially in today's tight economy. Although the e-business boom has turned into a bust, this e-commerce solution retains some leadership in executives' minds and technology initiatives. Cost cutting is rampant in the business world, and e-procurement embraces reduced product prices, enhanced productivity, streamlined processes, and increased profitability by

- Web-enabling the workforce,
- Establishing an electronic approval process,
- Expediting supplier relationships,
- Integrating supplier and buyer software systems, and
- Transforming an organization's operational culture.

An effective e-procurement solution will

- Decrease "maverick" buying,
- Reduce prices for materials and services,
- Reduce cycle time,
- Decrease administrative tasks and costs, and
- Support strategic procurement analysis and planning.

Fortune 1000 companies that have already implemented e-procurement solutions have clearly demonstrated the benefits and cost savings gained. The average transaction price of an order has been reduced 20 to 90%. Cycle times have been reduced 20 to 70%. The number of suppliers has decreased dramatically for all companies that have implemented an electronic solution. And of course, millions of dollars have been saved.

Hesitation and cautiousness are on executives' minds as well. Over the last three years, most companies have experienced reduced revenue, reduced profits, and economic problems. Layoffs have been abundant, bankruptcies in new technology companies are the norm. This has led to less spending on technology and slower decision making. The lesson learned from the e-business boom to corporate America is to analyze the companies with whom you partner and try to use technology and solutions that have already been proven. E-procurement is still in its infancy stage and will continue to grow as economic conditions improve. It will not be the boom of the 1990s, but e-procurement will continue to prove itself as a beneficial technology in savings, efficiency, information gathering, negotiation, control, and analysis. NatSteel Electronics has demonstrated four benefits of their e-procurement solution within the first year of implementation: reduced inventories, improved ability to react to customer schedule fluctuations, increased ease of doing business with the supply base, and less procurement staff. The benefits have just begun and will be proven even further over time.

GLOSSARY

Business-to-business (B2B) Transactions that take place between business entities, not between a consumer and a business.

Centralized procurement A single person or group of people responsible for all purchases within a company.

Decentralized procurement Purchasing throughout a company; purchases may be made by individuals, departments, offices, and so on. It does not require all purchase requests be sent to a single department within a company.

Direct procurement The purchasing of materials for the manufacturing of a product.

E-business A term used for all Internet- and electronic-related business, including Web sites, supply chain management, enterprise resource planning, e-procurement.

Electronic data interchange (EDI) A standard format for exchanging business data; the format is ANSI X 12. This form of electronic exchange is most often used in purchasing transactions and information.

Electronic procurement A technological purchasing solution that has greatly reduced or eliminated paper and manual processes. A business management system that integrates all facets of the business, including planning, manufacturing, finance, sales, marketing, purchasing.

Government-to-business (G2B) Transactions that take place between public and private companies and government entities.

Horizontal market Suppliers and products that focus on broad categories crossing multiple industries such as software, office supplies, and utility markets.

Indirect procurement The purchasing of materials not necessary for the manufacturing of products; purchase of materials for day to day activities and business.

Maintenance, repair, and operations (MRO) Activities that are directly related to the manufacturing of a company's product.

Operating resource management (ORM) Activities related to daily business but not directly associated with product manufacturing.

Supply chain management (SCM) Process-oriented, integrated approach to procuring, producing, and delivering products and services to customers.

Vertical market Suppliers and products that focus on a single industry, such as financial, telecommunications, or transportation and may cater to a core group of clientele.

Extensible markup language (XML) New standard developed by the World Wide Web consortium (WC3) to replace HTML (hypertext markup language) for development of interactive Web pages.

CROSS REFERENCES

See *Business-to-Business (B2B) Electronic Commerce; E-government; Electronic Commerce and Electronic Business; Electronic Data Interchange (EDI); Electronic Payment; E-marketplaces; Extensible Markup Language (XML); Supply Chain Management*.

REFERENCES

Aberdeen Group (2001, March 21). E-procurement: Finally ready for prime time. Retrieved April 10, 2002, from www.aberdeen.com

Aberdeen Group (2002, January 2). Aberdeen Group Newsletter. Retrieved May 1, 2002, from http://www.aberdeen.com

Ariba Solutions (2001). Retrieved February 10, 2002, from http://www.ariba.com

Center for Innovation, Business, and Manufacturing. (n.d.). *What is e-procurement?* Retrieved March 15, 2002, from http://www.cibm.sa.gov.au

Essig, M., & Ulli, A. (2001, Fall). Electronic procurement in supply chain management: An information economics-based analysis of electronic markets. J*ournal of Supply Chain Management, 37*(4), 43.

Giunipero, L. C., & Sawchuk, C. (2000). *E-purchasing plus.* Goshen, NY: JCG Enterprises.

Intel (2002). The e-procurement advantage. Retrieved May 1, 2002, from http://program.intel.com/solutions

Neef, D. (2001). *E-Procurement from strategy to implementation.* Upper Saddle River, NJ: Prentice-Hall.

Rybeck, D. (2001). Devil . . . or delight. Asia Associates Web site. Retrieved March 10, 2002, from http://www.asiassociates.com

Ageshin, E. A. (2001, Winter). E-procurement at work: A case study. *Production and Inventory Management Journal, 42,* 48.

FURTHER READING

Atkinson, W. (2000, September 21). Contract manufacturer uses e-procurement to become more competitive on costs. *Purchasing, 129*(5), S31.

Atkinson, W. (2000, September 21). E-commerce will cut stocks, shrink time-to-technology. *Purchasing, 129*(5), S68.

Carter, P. L. (1999). The future of purchasing and supply: Electronic commerce in the new millennium. Retrieved May 15, 2002, from http://www.ism.ws

Cisco/KPMG Consulting (2001). Solutions Brief. An eProcurement management solution. Retrieved May 15, 2002, http://www.kpmg.com

Cruz, M. (2000, May 8). E-procurement solutions enhanced developers cut costs, time from supply chain. *Computer Reseller News,* 51.

E-procurement (n.d.). Retrieved March 10, 2002, from http://searchhp.techtarget.com

E-procurement Initiative (2001). Retrieved April 5, 2002, from http://www.ucop.edu

E-procurement: The transformation of corporate purchasing (2000). Retrieved December 15, 2002, from http://www.fortune.com/sections/eprocurement2000

Ericson, J. (2001, August 23). What's next in e-procurement? Retrieved April 16, 2002, from http://www.line56.com

Foust, B., Shin, D., & Shehab, J. (2002). E-procurement apps hook onto the supply chain. Retrieved April 16, 2002, from http://www.informationweek.com

Greenemeier, L. (2000, April 3). Buying power: As online market places proliferate, businesses need to move fast to get e-procurement systems up and running. Software and services vendors can help speed the process. *Information Week,* 67.

Electronic procurement sets the stage for new marketplaces: Banks rush to adopt electronic-procurement platforms and convert them into business-to-business electronic marketplaces (2000, May). *Small Business Banker, 1*(4), 1.

Kasturi, R. (2000, September 8). Mission-critical e-procurement. *Intelligent Enterprise, 3*(14), 34.

KPMG Consulting (2001). Procurement transformation. Retrieved May 1, 2002, from http://www.kpmg.com

Marston, L., & Baisch, L. (2001, November). The overdue promise of e-procurement: A new business architecture and an evolutionary business model are necessary for electronic procurement to deliver and sustain the efficiencies that stakeholders require. *Health Management Technology, 22*(11), 32.

Moozakis, C. (2001, November 28). E-procurement gets priority. Retrieved May 15, 2002, from http://www.informationweek.com

Morris, A., Stahl, A., & Herbert, J. (2000). E-procurement, streamlining processes to maximize effectiveness. Retrieved April 10, 2002, from http://www.ebusinessforum.gr

Oscar, K. J. (2001, Spring). A common e-commerce architecture for the federal procurement system, *The Public Manager, 30,* 11.

Price, C. (2001, February 21). Bureaucracies resist e-market benefits: Electronic procurement in Europe, *The Financial Times,* 8.

Solutions for the best run e-businesses (2002). Retrieved February 10, 2002, from http://www.sap.com

Wendin, C. G. (2001, May 1). Slash purchasing costs. *Ziff Davis Smart Business for the New Economy,* 66.

E-mail and Instant Messaging

Jim Grubbs, *University of Illinois at Springfield*

INTRODUCTION

Although this chapter focuses on electronic or e-mail and the interactive equivalent known as instant messaging, neither technology would be as widely available as it is without the existence of an internetwork (Bucy, 2000). We know the largest of these networks as the Internet itself. For a more complete discussion, consult the entries for Internet History and Web Services: Past, Present, and Future in this encyclopedia. There really isn't a single Internet. Rather, there are a number of smaller networks that have been brought together in such a way that data routed from any point on any of these interconnected networks arrive at their destination in milliseconds. Many of the technologies and techniques discussed in this chapter existed in earlier forms long before the existence of what we now call the Internet. In those earlier days, the penetration and availability of electronic mail and instant messaging was severely limited—often to users all tied to the same computer.

The Internet has grown both in scope and in available bandwidth. Even well into the 1990s, in order to participate in an Internet-based videoconference, institutions had to shut down other communication temporarily to have available bandwidth to accommodate the video. Additional bandwidth and improved coding schemes now make those types of activities more common, if not routine. As bandwidth requirements go, individually, most e-mail and instant messages have a minimal impact on channel capacity. But as we will see, the sheer volume of such messages totals to a significant portion of network capacity.

Merging Technologies and Digital Convergence

It is becoming more and more difficult to clearly differentiate between communication devices. High-resolution display screens are used for both new television technologies and computer display purposes. The personal computer not only manipulates financial data and allows us to create text and graphic documents, but also allows us to listen to compact discs and other forms of digital music. We can watch DVDs on our PCs. The PC is capable of serving as a digital recording and editing studio. PCs now run and manage much of today's radio and television programming. DVD players can decode audio CDs as well as digital audio and video files encoded for PC use, which are not part of the DVD standard. PCs can be used to make voice phone calls using TCP/IP networks. Cellular phones can be used to exchange e-mail and browse the Web while also retaining their traditional role as a voice communication device. E-mail can be delivered to a pager; PCs and cell phones can send pages.

What makes all of this possible is that the majority of communication channels are rapidly being converted to digital data streams. Once data are in digital form, a CPU and associated circuitry can manipulate, encode, and decode them in nearly any form imaginable. Some high-resolution and high-bandwidth media require very fast processors and large amounts of memory, but the central concept remains the same. All of these media are moved from one location to another in digital form. The result is an unprecedented convergence of technologies.

Synchronous and Asynchronous Messaging

Modes of communication, whether electronic or otherwise, can be divided into two categories. When two or more individuals engage in real-time, interactive conversation, they are said to be communicating synchronously. That is, all parties have the ability to both talk and listen at the same time. The same term applies to synchronous communication among machines. For the purposes of this chapter, synchronous communication is represented by instant messaging. When we write someone a letter, or leave a message on an answering machine, we are communicating asynchronously. We send a message and trust that we will eventually receive a response, but the overall interaction is spread out over time. The Internet equivalent is electronic mail.

Each form of communication has both advantages and disadvantages. Synchronous communication is immediate. We can both ask and answer questions interactively in real or near-real time. Asynchronous communication allows us contact with others at a time of our choosing.

Impact on Society

In 2001, e-mail, in the general form that we know it today, celebrated its 30th birthday. International Data Corporation estimates that 9.8 billion messages are sent daily with the number of messages continuing to rise (Hafner, 2001). Compare that to approximately 300 million pieces of e-mail as recently as 1996 (Hafner, 2001). It is important to note that 1996 was also the year at which the number of e-mail messages first equaled and then surpassed the number of pieces of postal mail sent on a daily basis. Long before the public popularity of e-mail, there were early indicators of its importance as a communication medium. In the early 1970s, more than 75% of the data exchanged on ARPANET were e-mail, even though that was not the task for which the system was designed (Hafner, 2001). For a quick overview of e-mail statistics and milestones, see Table 1.

In the corporate world, e-mail is the new interoffice memo. E-mail is asynchronous; users send and receive messages at their convenience. E-mail also allows a user to send the same message to a group of people with ease. Though text-based messaging is very inexpensive, predictions that e-mail would replace existing forms of communication were only partially correct. There is evidence that e-mail has created entirely new levels of communication in addition to the existing levels that it enhances or has supplanted. Twenty years ago, governments predicted that there would be a significant reduction in postal workers as the result of widespread use of e-mail, but in fact, the volume of traditional mail has doubled during that time; there has also been a significant increase in postal service employment rather than a decline.

Table 1 E-mail Milestones

Year	Milestone
1971	First e-mail sent
Circa 1973	75% of ARPANET traffic was e-mail
1981	General Accounting Office predicts postal employment would be cut by two-thirds by 2000 as the result of e-mail
1982	Postal Service E-com Mail introduced (electronic computer-originated)
1983	MCI mail introduced
1985	E-com mail product discontinued by postal service
1996	300 million e-mail messages per day (average)
2001	9.8 billion e-mail messages sent each day

Source: (Hafner, 2001)

Early digital communication modes, including the Morse telegraph, long ago broke down barriers of time and space. Although most of us associate the sounds of "dots and dashes" with Morse code, the earliest form relied on the "clicking" of a simple sounder, while the spacing between the clicks allowed for decoding. The instant messaging possible via electronic communication had a number of effects on society. Standard time, for example, replaced "sun time" so that merchants and traders could conduct business without confusion; the telegraph helped trains to run on a regular schedule. In the 20th century, e-mail has had an equally profound effect on society; documents of all kinds navigate around the world through the Internet in small fractions of seconds, and without any loss in detail (Rheingold, 2000). Time, and even cost, have all but been eliminated as a factor in communication, but these advances have not come without some potentially negative effects as well.

E-mail offers a way to remain relatively anonymous, but at the same time has provided a means for words to be easily archived and retrieved years after a message is posted. Unscrupulous individuals can exploit e-mail attachments to infect distant computers with a virus or other program designed to destroy data on a personal computer or even a corporate server. It is very easy to instantly send a message to the wrong person or group, with no way to retrieve the message once the send function starts. Users are besieged with unwanted or spam messages because e-mail distribution makes it easy to distribute such communication.

Marshall McLuhan (in McLuhan & Zingrone, 1995) noted that new media at first mimic older media. E-mail, in order to gain wide public acceptance, advanced from command line, text-only messaging with few features to a form inspired by more traditional written communication. Today, we see e-mail as a versatile form of communication, but the earliest e-mail programs did not even include a "reply" function. Once e-mail achieved a look and feel similar to traditional communication, it started to explore new and innovative ways of communicating. Mail with graphics and publication-style layouts is now widely used, thanks to the inclusion of HTML parsing in e-mail clients. It is now possible to include audio and video elements in e-mail in addition to text-based messages.

In addition to asynchronous, "send and receive messages at your convenience" forms of electronic communication, a parallel application exists for synchronous or "real time" communication. Popular instant messaging programs such as America Online's Instant Messenger

(AIM), Microsoft Messenger, and Internet Relay Chat (IRC) clients allow two or more users to communicate. These applications are fast becoming multimedia capable. Even Napster and other file-sharing applications are, at their core, based on IRC-type communication.

As we have witnessed on multiple occasions, our reliance on and penchant for e-mail can cause problems when e-mail systems fail to operate as designed. For example, in the 1990s, America Online frustrated large numbers of users when their infrastructure became unable to keep up with the load demanded by their customers; smaller but equally disruptive network outages occur on a daily basis. When business depends on electronic communication, it is hard, if not impossible, to conduct business when the technology is not working.

The Impact of E-mail on Terrestrial Mail and Facsimile

Although the public was slow to embrace e-mail during the 1970s and 1980s, facsimile or fax communication grew rapidly. Facsimile technology was not new, but the means to produce machines economically made the medium attractive to a wide range of users, especially businesses. Ironically, after the introduction of the personal computer in the 1980s, businesses regularly created documents in electronic form, only to print them and then fax the printed documents to distant points. On the receiving end, if an electronic copy was required, the document had to be rekeyed. Users were slow to realize that the file could be sent directly, without the intermediate fax step.

Today's messaging technologies make it easy to integrate a variety of communication channels. Even when the recipient prefers a fax transmission, it is now an easy matter to start the document as an e-mail message with automatic translation and delivery via a fax server. Speech synthesis and speech recognition are making it possible to both send and receive e-mail messages using telephone circuits.

How Instant Messaging Has Changed Telephone and Face-to-Face Communication

Instant messaging has had multiple effects on line usage time. It is just one of several new communication modes that have changed the way we talk with each other. The rapid growth in Internet usage has actually had a positive effect on line usage time. It is more difficult to identify the percentage of that increase attributable to voice communication. Logically, much of the increase can be attributed to modem use. Demand for second and even third lines into private residences has skyrocketed as more and more households connect to the Internet. Cellular telephones and other mobile/wireless communication devices now supplement fixed-location phone service. The result has been a shortage of available telephone numbers, the splitting of area codes to create more available numbers, and more complex dialing requirements in order to accurately route traffic to more and more numbers. In 1997, Regional Bell Operating Companies (RBOCs) noted their concerns about Internet usage resulting in network congestion. Cable modems, some satellite connections, and DSL service provide full-time Internet connections without tying up the existing phone line or requiring additional phone numbers to be assigned. Additionally, home networks allow multiple computers to share a single Internet connection, further reducing the need for additional telephone lines.

As for instant messaging, there are strong indications that it has had a significant impact on the way young people, especially teenagers, communicate. America Online report that their system services more than 430 million instant messages a day (AOL Signs Messaging Allies, 2002). In 2001, the Pew Internet and American Life Project noted that three out of four online teens use instant messaging (Pew Internet and American Life, 2001). Twenty percent of these teens rate instant messaging as their primary means of communication with friends. More than a third use instant messaging to say something they would be unlikely to say by phone or in person. It is fashionable for teens to give others their screen names rather than their phone numbers. The topics of conversation are not new, but the method of communication is. Teachers also use instant messaging to stay in touch with students and to offer schoolwork help.

Instant messaging allows multiple, simultaneous, individual conversations. The Pew study found teens typically carrying on three conversations at once, known as "split attention." Researchers of online behavior question whether adults have developed or can develop the ability with the same ease as teens do. The Pew study found that less than half of adults active on the Internet have ever tried instant messaging.

Even the classroom is not immune to instant messaging. Using wireless devices, students send instant messages to friends while in class. The study found that students engaged in such behavior, along with college students who used instant messaging late into the night, got lower grades and missed more classes. Many online adults have also adapted to the instant messaging medium to keep in touch with family or, in some cases, to look for companionship. Where it was once common to make a weekly phone call to distant family, almost daily instant messaging sessions have replaced many of those conversations. Business people use instant messaging to stay in contact with team members. For example, an engineer working on a project at a customer's location can consult with other engineers at distant locations without trying to balance a telephone handset at the same time. Other businesses are offering online help and support via instant message.

Instant messaging also provides a tool for the deaf community. In the 1960s, an amateur radio operator created a program to recycle older models of teletypewriter machines (TTY) as communication devices for the deaf (Grubbs, 1988). Advances in solid-state technology allowed the creation of smaller telecommunication devices for the deaf (TDD). Such machines generally are self-contained and connect via specially designed acoustic modems to standard telephone handsets. TTY and TDD machines cannot directly connect to the Internet. As late as the 1980s, a significant portion of the deaf community resisted the switch to the text communication possibilities of the home computer. Much of the problem was one of compatibility. PCs were still a novelty, whereas the TTY/TDD network was well established. With the proliferation of PCs and the introduction of

AOL Instant Messenger in the 1995, the deaf community quickly embraced the technology. Instant messaging serves as a tool which levels the playing field for the hearing-impaired. Online, a deaf individual is perceived the same as any other user; interpreters are not required, and meetings can occur with both hearing and hearing-impaired participants without special arrangements. The types of communication most people have taken for granted are now convenient for the hearing-impaired as well. The one remaining obstacle has also recently been shattered. For example, with devices like the BlackBerry (Blackberry, 2002), e-mail and instant messaging are available in a mobile, wireless environment at a reasonable cost. The BlackBerry is a wireless handheld device with optional data and phone services. The hardware itself is less than $400—the cost of a full-featured TDD device—with an additional monthly service charge in the $20 range.

THE HISTORY OF E-MAIL

The First E-mail

Whereas "talk" or chat commands allowed synchronous communication among computer users, early pioneers saw the need for and usefulness of an asynchronous method of communicating among themselves. Ray Tomlinson, of Bolt, Beranek, and Newman, is credited with sending the first e-mail in 1971. Using the command-line programs SNDMSG and READMAIL, Tomlinson sent the first message to himself—to prove the concept. E-mail was soon adopted by his colleagues as a part of their regular communications. A number of e-mail servers and clients were soon developed.

BITNET Mail

BITNET was a network specifically designed for the communication of electronic mail. It ceased operation on December 31, 1996 (CREN History Winding Down, 2002). Though it is now primarily of historic interest, much of today's electronic messaging can be traced to the days before the Internet existed and specifically to BITNET. It began in 1981, providing the City University of New York and Yale University with a way to exchange messages. It grew to more than 500 institutional members in the United States alone. Internationally, BITNET connected with 1,400 institutions in 49 countries (CREN History and Future, 2002).

The routing and propagation of e-mail on BITNET used a concept different from that common on the Internet. Rather than an end-to-end connection being established between two distant points for direct transmission of messages between mail servers, the "store and forward" method used interim nodes connected by leased lines. In some cases, BITNET connected systems that relied on dial-up connections. Although it was based on an IBM standard, VAX, UNIX, CDC, and other computer systems could be used as long as they were capable of implementing the IBM NJE (Network Job Entry), a "store and forward" protocol.

Movement of messages from one node to the next was usually accomplished by regularly scheduled transmission and reception of messages grouped together for bulk delivery between major nodes. Especially in the early days of the system, messages could take days, even weeks, to arrive at their destination. Although BITNET itself has been supplanted by other communication means, one of the elements of the system lives on in a very impressive fashion. For ease in communicating with groups of people, LISTSERV, a mailing list server, is still widely used. Over the years, approximately 3,000 lists, covering most academic areas, have been created. As many as a thousand or more participants per list contribute to the discussions.

Today's Internet could be used for BITNET communication, but BITNET was an independent network designed to be self-sufficient. For a time, gateways allowed the exchange of electronic mail between BITNET and the Internet (NSFNET and its associated regional, state, and campus networks) and other networks.

Usenet/Newsgroups

The "User's Network" or Usenet is commonly referred to as the "newsgroups." Thousands of virtual bulletin boards are dedicated to just about every imaginable topic. There are currently eight hierarchical categories of discussion groups. See Table 2 for a list and explanation of each category.

A ninth category, called "alt," exists outside the officially defined eight general topic areas, even though it encompasses more newsgroups than any other category. It exists as a way to ensure free speech on the Usenet. In order to create a group in any of the eight main topic areas, a specific procedure must be followed. This is not the case with "alt" groups; anyone can create one. Usenet newsgroups can be particularly helpful in searching for answers to questions. An interface, such as that provided at http://www.google.com, allows anyone to search an archived database of newsgroup postings.

Newsgroup messages are created much like regular e-mail messages, though they propagate through the net in a somewhat different fashion. Network News Transfer Protocol (NNTP), developed at the University of California at Berkley, is the most common propagation mechanism, but others are used as well (Internet FAQ Archive, 2003).

Table 2 Usenet/Newsgroup Hierarchy

Usenet Hierarchy Name	Subject Matter
COM	Computer science subjects
HUMANITIES	Humanities subjects
MISC	Miscellaneous groups
NEWS	News topics
REC	Recreational subjects
SCI	Science topics
SOC	Sociological subjects
TALK	Controversial topics
ALT*	Alternative–any topic imaginable

*Not part of the official Usenet hierarchy.

Commercial E-mail

In 1983, MCI (now WorldCom) and CompuServe (now part of AOL Networks) provided the first commercial e-mail service (Hafner, 2001). A 500-character message cost 45 cents to send. Though messages were delivered electronically almost immediately, clients were not accustomed to checking for electronic mail. MCI offered a service whereby MCI employees would telephone recipients to let them know they had electronic mail and encourage them to check for it. From 1982 to 1985, the U.S. Postal Service offered a service known as electronic computer-originated mail (E-Com). Corporate clients could create messages electronically for delivery over a postal-service-controlled network (Hafner, 2001). If the recipient was not also an E-Com client, the message traveled electronically to the distant post office, where it was printed in hard copy form and delivered via standard postal mail. The advantage was that virtually all of the long-distance travel time was eliminated. This type of communication really was not new to the business world. Previously, telex messages flowed between major corporate clients. Telex machines were, in essence, teletypewriter machines connected via leased telephone circuits. Unlike e-mail, the telex machine was a dedicated device; when a message came in, the paper fed through the machine, making it obvious that one or more messages had arrived.

THE TECHNOLOGY BEHIND E-MAIL

Addressing, Routing, and Sending E-mail

The form of e-mail address familiar today dates back to a convention developed by Tomlinson for his e-mail system. He wanted a way to separate the user's name from the name of the computer to which the user was connected. The @ symbol became the separator (Hafner, 2001).

For e-mail to travel efficiently and effectively from one user to another, a central database of routing information is required. For further information, see the entries in this encyclopedia on transmission control protocol/Internet protocol (TCP/IP), digital communications, standards and protocols in data communication, Internet architecture and protocols, and public networks. E-mail is a subset of Internet traffic that relies on the TCP/IP standard. Those parts of the routing process unique to e-mail are discussed here. Internet addressing (and therefore, routing) is based on the assignment of unique IP addresses. It is possible to transmit and receive data of any sort, including e-mail, by knowing the IP address of the receiver. Early network developers soon realized that it is contrary to human nature to remember large groups of numbers. As a result, plain text "domain names" were introduced. For example, "uis.edu" is the domain name for the University of Illinois at Springfield. A domain name is independent of the route used to reach the domain.

Each domain has a set of records associated with it that help to route communication between two points on the network. One of the records is the mail or MX record. The mail server for the domain is specified in this record. Even in cases where mail is handled by a server other than the one used for other Internet communication, mail addressed to the domain is properly routed to the correct mail server. Following the convention, an e-mail address looks like this:

username@hostname.domain

The actual protocol used to exchange mail between mail servers is called the simple mail transfer protocol or SMTP (ISOC Internet Standards Programs, 2002). The server software used varies among operating systems. In Unix, sendmail is the most common SMTP server used. The SMTP protocol defines a set of rules for the exchange of messages. For example, SMTP communication begins with the HELO (hello) command; the actual exchange of the message begins with the MAIL command and ends with a QUIT command. These rules also include a method for determining the validity of the user name. Once the message has been exchanged, the received message is stored on the receiving machine's mail server until it is accessed by the user to whom it is addressed.

Receiving E-mail Using POP3

Post office protocol or POP (ISOC Internet Standards Programs, 2002) is a set of rules used by e-mail clients to both send and receive electronic messages. Its third release, POP3, is one of the most commonly used protocols for Internet e-mail. POP3 e-mail accounts are protected through the use and verification of account names and passwords. Various authentication systems are used to maintain password security. Once incoming messages have been read, they can either be deleted from the server or kept on the server. The local e-mail client can recognize previously received messages (because they have been flagged as read) and will not download them again. Passwords themselves can be encrypted and stored using DES (data encryption standard) or other schemes for protection. See the chapter on passwords in this encyclopedia for further information.

Receiving E-mail Using IMAP4

The Internet message access protocol (IMAP), developed at Stanford University in 1986, is a newer method of e-mail exchange than POP3. Like POP, IMAP is being continuously updated and is now in version 4 (ISOC Internet Standards Programs, 2002). IMAP4 includes features not found in POP3, including an enhanced "store e-mail on server" function. This feature provides functionality for users who travel, for example, and who prefer to keep all their mail on the server. It is also possible to get a list of pending messages but then only download a subset of those messages. This is especially useful when limited bandwidth is available, as is often the case with PDAs or e-mail-enabled cellular telephones.

HTML in E-mail

With the exception of e-mail attachments, which can take virtually any form, the body of an e-mail message is alphanumeric in nature. Whether one of the variations of the ASCII character set, or a specialized character set for another language, the core content of an e-mail message is alphanumeric. In an attempt to make e-mail more attractive, software developers realized that HTML code could be used to create e-mail with images and publication-type formatting. HTML is itself a text-based markup or formatting language. No special software is required to send a message with HTML in it. The situation is different on the receiving end. In the absence of an HTML interpreter, the message will appear as a plain text message that includes the HTML tags. In essence, e-mail clients that properly display HTML-encoded messages are using a simple HTML browser to interpret the HTML code and display the message as the sender intended.

On the sending end, e-mail software that allows a user to create an HTML message includes a simple HTML editor. All of this is generally transparent to the user.

Some e-mail clients implement only a small number of HTML tags. In that case, a message may properly display fonts and simple formatting commands, but be incapable of integrating graphics and more complex layouts. Put in the simplest terms, HTML-enhanced e-mail is a Web page contained in the body of an e-mail message. That is not the same thing as sending an HTML file as an attachment.

E-mail Attachments

Especially in the early days of networking, the channels used for exchange of e-mail consisted of circuits designed to exchange plain text messages. That presented a problem when binary files not encoded in plain text were exchanged. The solution was the creation of the Multipurpose Internet Mail Extensions (MIME) standard. MIME encoding of attachments is widely supported across the Internet and in e-mail clients. Regardless of the content of the attachment, MIME encoding changes everything from pure binary code to 64 alphanumeric characters:

ABCDEFGHIJKLMNOPQRSTUVWXYZ

abdefghijklmnopqrstuvwxyz0123456789 + /

The binary bits are encoded and translated into a hexadecimal code that can be represented by the characters shown above. On the receiving end, the hexadecimal-encoded text is translated back to binary. Additional instructions in the alphanumeric encoded text tell the decoder what MIME "type" the attachment is. For example, JPEG and GIF pictures have their own mime type, MPEG audio and video files have their own mime type, and so on.

Voice and Video in E-mail

Including voice or video in an e-mail message is as simple as attaching a properly encoded file containing the voice or video message. Although simple in theory, the challenge is to create a file of modest size so that it does not exceed the attachment limit on some e-mail servers. The solution to the problem is to use an encoding scheme that heavily compresses, or reduces, the total file size through a controlled reduction of redundant code. An example is Qualcomm's "Purevoice" technology. Internet e-mail users associate Qualcomm with the Eudora mail client, but their primary business is CDMA wireless (cell phone) technology. The result is a voice data stream that is very compact and requires very little bandwidth. Fidelity is limited to that of a standard voice connection.

There is another way to overcome transmission problems. Services such as AT&T WorldNet allow users to send video e-mail, but the video portion of the e-mail is never sent to the recipient. Rather, the video is stored on a video server. The text portion of the e-mail includes a URL to allow the recipient to access a streaming version of the file.

INSTANT MESSAGING

Unix "Talk"

Synchronous, "real time" communication of personal messages between computer users has been around since the first computer network was created. In the absence of an Internet or even an intranet, early implementations required that all users be hooked up to the same computer. Users accessed the central computer using "dumb" terminals. Terminals generally do not have any computing power. They are merely display and input devices connected to a central computer via hard wiring. With the development of UNIX by Bell Laboratories, the "talk" command became something of a standard for instant messaging. It was used extensively during the 1970s and beyond as a means for synchronous communication over ARPANET. In the UNIX environment, everything was done at a command line prompt. There was no graphical user interface (GUI). The talk program in UNIX was not very sophisticated. Messages from other users appeared on a terminal, interrupting whatever task the user was currently working on.

PLATO "Talk" and "Talkomatic"

The PLATO system was created on the Champaign–Urbana campus of the University of Illinois in the 1960s under the auspices of the Computer-based Education Research Laboratory (CERL). PLATO was designed specifically for computer-based education. Over time, a number of communication options were designed as part of the system. Unlike UNIX, the system included high-resolution touch-screen, graphic terminals that allowed complex graphics to appear on the screen rather than only standard ASCII characters. Talkomatic was the original chat client for PLATO. Characters were transmitted in real time as they were typed. Other talk programs generally delayed transmission until a return character was received from the keyboard. Several horizontal windows appeared on the screen—one for each participant in the chat.

Over time, the program was adapted to allow multiple communication channels, with each channel capable of handling five interactive users. The system allowed an unlimited number of users to monitor a particular channel, though an active user had the ability to "protect" the channel and allow only the active participants to see what was being exchanged. Talkomatic was constantly improved to the point of including many of the features we associate today with Internet-based instant messaging programs.

Internet Relay Chat

Internet Relay Chat or IRC is distributed, real-time communication. IRC traces its roots to the early UNIX Talk command and the associated Talker program. Talker programs are generally proprietary and run on single servers with a maximum of several hundred simultaneous users. Conversely, IRC is a multiserver, networked system capable of allowing thousands of users to use the system concurrently. IRC is an open Internet protocol invented in 1988 by a Finnish graduate student, Jarkko Oikarinen. Inspired by BITNET's Relay Chat and other existing programs, IRC became a reality in August 1988. Shortly after its introduction on Oikarinen's local campus, he asked friends at other Scandinavian universities to run IRC servers to help distribute the load. Soon after, MIT, the University of Denver, and Oregon State University joined the distributed server network As detailed later in this chapter, IRC got a boost in public awareness following the 1991 invasion by Iraq of Kuwait.

When a user starts an IRC client, he or she is connected to servers that distribute the messages. The purpose of the server is to echo the communication on a number of IRC channels. Although none of the users are directly connected, the IRC server creates the illusion that they are. Messages sent by each user are relayed to all users on the channel very quickly. By connecting distant IRC servers together over a network, the channels available on each server can be made available to all users on all active servers without the user having to take any special action. By sharing the load, both the number of channels and the number of users can grow without too great a burden on any single server. The distributed networks of servers share all of the traffic with each other. On a very busy network, there may be a delay of several seconds for messages coming from a distant server. More than a half dozen large IRC networks run simultaneously, each with dozens if not hundreds of servers. Tens of thousands of users can be accommodated at any given time.

In IRC language, a channel is a virtual meeting room. Users commonly use the term "chat rooms" interchangeably with channels. Users enter, participate, and leave channels continuously. Some client software allows a user to participate in multiple discussions on different IRC channels at the same time. Channels for just about every discussion topic and interest imaginable exist. When someone comes up with an idea for a new topic, it is a simple matter for that user to create a new channel. IRC channels are to instant messaging what LISTSERV is to e-mail.

IRC During Historic Events

During much of the 19th and 20th centuries, we relied on mass media to provide news on world events. At first, the daily newspaper was the means for disseminating dispatches from around the world. There were inherent delays. Even when events local to the newspaper were involved, there was a significant delay while a special edition of the paper was created and distributed. The advent of the telephone provided the means for virtually instant communication, but well into the 20th century it remained available only to the wealthy. The widespread availability of radio in the 1920s and 1930s provided a means for communicating information to the masses in a timely fashion. However, the content of those broadcasts was controlled by a relatively small number of individuals.

Radio amateurs ("ham operators") provided ad hoc communication networks as needed during times of disaster (Grubbs, 1988). Later citizens' band (CB) supplemented the amateur radio ranks with a large number of radio-equipped private citizens. These men and women still provide communication in disaster situations, but cellular and other wireless technologies have provided an alternative means of communication during such events. In the 1970s and beyond, the fax machine became a means of distributing information from a single source to many others in a very short period of time. The 1989 events of Tiennamin Square in China illustrated just how effective private fax machines could be at distributing information from "unofficial" channels around the world. Had the event happened just a few years later, the same information might have traveled worldwide via e-mail and IRC channels.

By 1991, the Internet was well enough established so that it began to rise as a favorite communication channel during difficult times. The 1991 invasion of Kuwait brought IRC to the public's attention. Even though traditional media outlets had been cut off, an IRC link was still operating a week later, with the result an exponential growth in IRC logins. Another event in 1993 resulted in a similar interest in IRC. As Russian lawmakers sealed themselves in Parliament, both Russian and American users relied on a specially created IRC channel to disseminate information around the world. IRC has been used in natural disasters as well. For example, when the 1994 Northridge earthquake in California knocked out many communication systems, Los Angeles Internet users established a channel to communicate with others outside the quake area (Why IRC Is Important, 2002).

Although we may never know the full extent to which Internet communication was used both in planning and in reacting to the September 11, 2001 attacks, we do know the tremendous amount of Internet traffic generated in the period shortly thereafter. IRC channels buzzed with news of the events worldwide. In all of these cases, this communication sprang up spontaneously, and for the most part, beyond government control. These events provide us with an indication of the power of the Internet itself and communication tools such as IRC.

AOL Instant Messaging

In addition to IRC servers run for the public, other proprietary chat networks exist. Perhaps the most notable is the instant messaging network available through America Online. The large user base for AOL has helped AIM to prosper. AOL subscribers automatically get the use of AOL's AIM instant messaging software. One of the unique features of AIM is that it can also be used by non-AOL subscribers. Newer versions of the Netscape browser also include AIM as part of the installation package. Interaction between members and nonmembers is seamless, as long as the nonsubscriber uses AOL's own instant messaging software. AOL members and nonmembers alike can access AOL's proprietary messaging system via a

Java-enabled Web page. The underlying software running AIM is similar to, but not compatible with, IRC standards. Early versions of the AOL instant messaging software did not allow group chats; the latest versions include group communication.

ICQ ("I Seek You")

The first release of ICQ came in November 1996. The design of the software embraced a new philosophy for locating and alerting online users for instant messaging. Current versions of the software allow not only instant messaging and chats, but also e-mail, SMS, and wireless-pager messages. File transfers and voice connections are also possible with ICQ. Although other messaging clients may be better known, ICQ has long been an innovator (*ICQ*, 2003). For additional information on audio and video transmission on the Internet, see other entries in this publication for related topics.

Microsoft Messenger and Other Instant Messaging Programs

Microsoft Messenger is that company's entry into IRC-like communication. Messenger comes packaged with an e-mail account (Hotmail) and other Microsoft products. Barnako (2002) reports that in May 2002, MSN Messenger was used by 15% of Web users, whereas AOL's Instant Messenger was used by 21% of Internet users. Like AOL's AIM, the underlying functions mimic IRC but are based on proprietary code. Yahoo Web Chat and Google Chat, as well as a number of private party software products, implement real-time chatting through Java-based clients accessed through a Web browser. This type of implementation tends to be significantly slower at distributing messages to chat members.

Interoperability Issues

Regardless of what IRC client is used, it can generally be used on any chat system that conforms to IRC published standards. Because many clients are written by hobbyists, there can be bugs. Unfortunately for the end user, proprietary systems such as AOL's AIM cannot be accessed via a standard IRC client. Similarly, an AIM user cannot directly connect with a Microsoft Messenger user. The only workaround is to run multiple chat clients simultaneously. Several software developers have developed multiprotocol clients that allow users to connect and interface with a number of different chat services through a single graphical user interface. For users of the Windows operating system, a program called Trillian allows access to Microsoft Messenger, Yahoo Messenger, and IRC using a single interface (Trillian, 2002). In the UNIX world, EveryBuddy is an example of such a client.

FUTURE TECHNOLOGIES

Although a number of messaging technologies have been introduced, particularly in the area of wireless and mobile applications, mere existence has not equaled acceptance (Grant, 2002). A 2001 study by the Yankee Group revealed that less than half those who tried wireless Web devices became regular users. The same group pegged wireless phone Web usage in the United States at the end of 2000 at 4.4 million users. In 2002, approximately half of wireless users had access to Short Message Service (SMS). Current predictions are for wireless Web usage to increase to more than 56 million users by 2005. In Europe, approximately 10 million users had access to the wireless Web in 2001 (Garcia, 2001).

The Potential of Wireless and Mobile E-mail and Instant Messaging

In the 1980s, amateur radio operators developed a form of wireless technology based on packet networking schemes to exchange e-mail messages. Admittedly, mobile operations were cumbersome, but they were possible. The same store and forward packet-based networks still operate on small LEO (low earth orbit) communication satellites and even the space shuttle. Commercially available mobile modems of the 1980s allowed laptops to connect to traditional dial-up services through mobile phone connections. The system worked, but was bulky.

In September 1999, Sprint PCS began offering wireless Web service aimed at the general consumer market. A product targeted for business users was introduced a year later. Today, virtually all mobile carriers offer some type of Web- or e-mail-like messaging service.

There are two primary obstacles to the success of devices that incorporate these features. As currently configured, data rates over wireless cellular networks are 14.4 kbps. Users accustomed to even a 56K modem find the resulting load times extremely long. Until additional bandwidth is available, there is very little providers can do to make access faster for wireless and mobile users. GSM (Global System for Mobile Communication) has the potential for bit rates up to an over-the-air bit rate of 270 kbps, although speech is still encoded at 13kbps. First launched in Japan in February 1999, NTT (Nippon Telephone and Telegraph) has developed I-mode technology—a competitor to the Wireless Application Protocol (WAP). I-mode is limited to a subset of HTML. As a result, I-mode pages do not require the type of translation server necessary with WAP. I-mode is capable of full-color presentation, is always on, and is packet-switched. Parallel voice and data networks make the "always on" feature possible. Normal HTML pages do need to be recoded with I-mode limits in mind, but the translation required is minor in comparison to that required for WAP server operation.

NTT markets the technology under the name DoCoMo—in Japanese, doco mo means "anyplace you go." In English, the acronym stands for "Do Communication Over the Mobile Network." Initially, data transfer speeds were relatively slow; however, with the introduction of a third generation (3G) network based on W-CDMA (wideband code division multiple access) technology, data rates up to 2 Mbps are possible. U.S. interests are developing and deploying GPRS (general packet radio service). GPRS is designed for data rates up to 115 kilobits per second, though it can be configured for a number of different bandwidths to accommodate a variety of services and situations.

The second challenge is associated with the input and output devices available on the typical wireless device.

Keyboards either are very tiny or require special manipulation of a numeric keypad in order to create a full range of alphanumeric characters. There is some hope for improvement here. Improving voice recognition technology can eliminate the need for routine use of a keyboard.

A 2-inch-square display screen is about average for most cell phones and other wireless devices; generally, the display is also monochromatic. Although this works well for short text messages, it is not a very effective output device for general Web content. The problem is similar to that faced by users of products such as WebTV, only worse. That is, in order to navigate most Web pages, it is necessary to scroll both horizontally and vertically in order to view the area normally available on even a small, low-resolution monitor. Wireless users are best served when the Web content is specifically formatted for a very small screen. As other electronic ventures have learned, a technology can be driven by available content; without content, the technology fails to find a home with the consumer.

The Wireless Application Protocol (WAP) allows HTML-coded Web pages to be converted on the fly to text only. All of the graphics and much of the special formatting are lost in the process. WAP does not support color. Carriers have tried to make the wireless Web somewhat more attractive by including menus of commonly used Web services known to be most compatible with WAP. Accessing sites not on the menu is still possible, but requires entering a URL using the small keypad on the device. Wireless devices also have limited memory and less computing power than larger devices, further limiting their ability to cope with richly featured Web pages. Single-purpose devices that allow both sending and receiving e-mail are also available. Research In Motion Limited of Canada has experienced success with a device known as BlackBerry (Blackberry, 2002). As of April 2001, Research In Motion had sold more than 200,000 units (Is Wireless Instant Messaging the Future of Communication? 2001). The unit includes a contact list (address book), calendar, and task list. Handspring offers devices that integrate cell phone and PDA technology with wireless Internet access. Treo and VisorPhone are two such products.

Successes in Europe, Japan, and Southeast Asia

The European and Japanese markets (as well as other areas such as the Philippines and Malaysia) have reacted more favorably to wireless and mobile applications. Carriers are aided in their attempts to attract consumers by the existence of a single transmission standard. The American open market allows competing technologies and standards. Although competition often results in new and innovative techniques, it makes it more difficult to capture the level of market share required to mass-produce devices with features and prices attractive to consumers.

Perhaps the most successful form of wireless instant messaging is a communication form well suited to the limited bandwidth and small display size (but still challenged by input device problems). In both European and Asian markets, short text messaging, more properly known as Short Message Service (SMS), has been popular for several years. SMS allows messages limited to 160 characters or less to be sent between SMS-equipped devices. Teenagers and young adults seem to have adopted the medium as their own. In some cases, SMS is only possible among customers who subscribe to the same system. Even so, in 2001 the GSM Association estimated that, each month, on an international basis, over 10 billion SMS messages were sent (Is Wireless Instant Messaging the Future of Communication? 2001). The *Wall Street Journal* reported that in May 2002, the number of SMS messages had risen to 24 billion per month (Taylor, 2002). In some areas, an SMS number has been established by police for deaf and mute persons, enabling them to take advantage of the safety of wireless communication (Taylor, 2002).

Presence Awareness

Instant messaging allows us to be aware of the presence of another person on the network. Currently, our awareness is limited in two ways. First, instant messaging software is generally a separate application; it is not integrated into the other computer tools that we use everyday. Second, we currently restrict our "presence awareness" technologies to person-to-person messaging. The future will include native instant messaging capability built directly into applications such as word processing programs, spreadsheet programs, and just about any enterprise where collaboration is useful. The possibility for generating new revenue streams exists. For example, an instant messaging client integrated into tax preparation software would allow revenue from tax preparation advice.

Today, collaborators can use a shared whiteboard product such as Microsoft's NetMeeting, but direct person-to-application messaging or application-to-person messaging can facilitate the process directly from the main application. Even machine-to-machine communication, for applications with "presence awareness," is possible.

"Session Initiation Protocol" (SIP) and "SIP for Instant Messaging and Presence Leveraging Extensions" (SIMPLE) are the protocols behind the idea (SIP Center, 2002). Interoperability (designing applications to the SIMPLE protocol) will be a key factor in the development of this technology. Lotus already incorporates "Sametime" instant messaging based on the standard.

E-mail from Your Toaster

E-mail and instant messages need not be limited to human communication or even communication between people and applications. Using "message queuing" (MQ), devices of all kinds can communicate with each other. IBM's MQSeries products, for example, allow applications on a variety of platforms to communicate with each other. The only requirement is an agreement on the content of the message (Kruczek, 1996). Communication-enabled appliances can be controlled by e-mail, issue reports on their status via e-mail or instant messages, or even communicate with each other.

Issues for E-mail and Instant Messaging

One of the challenges facing both the industry and end users is integrating convenient and accurate encryption into the client software we already use. Encryption methods such as "Pretty Good Privacy" (PGP) are available and work well, but tend to be cumbersome for users to implement. Although e-mail handles multimedia materials fairly well using the MIME format, some applications, such as Microsoft Outlook, have made alterations to the format rendering other e-mail clients unable to receive attachments under some circumstances. The areas of speech recognition and language translation also have the potential to make e-mail even more valuable. At the present time, reliable speech recognition remains elusive, and language translation, although possible, is still under development.

What's On the Horizon?

Gordon Moore, one of the founders of Intel, made an interesting observation in 1965. By plotting the capacity of microchips, he noticed an unmistakable trend: capacity was doubling every 18 to 24 months. That is, the amount of information storable on a given amount of silicon (a microchip for example) will double during the period. We now know this observation as Moore's Law. It rates escalation to a law because over the last 40 years it has been proven true over and over again. A corollary to Moore's Law is that as capacity doubles, prices halve during the same period.

George Cox, director-general of the Institute of Directors in the United Kingdom, notes that whereas it is relatively easy to predict the capacity of the technology itself, what is much less predictable is the social effects of technology. So although we can imagine and design communication devices incorporating artificial intelligence and virtual reality, what that will mean to society is nothing more than a guess.

In the immediate future, a number of technologies are ripe for deployment once the average user has access to enough bandwidth to accommodate the communication medium. We continue to see strides in coding techniques that allow many hours of full-motion video to be stored in devices that previously had trouble dealing with more than a modest number of textual data. Voice and video mail over the Internet are two technologies that will benefit from additional bandwidth. With enough available bandwidth, text-based messaging may become the exception rather than the rule.

In the wireless and mobile environment, issues of input and output will regulate growth. Voice recognition is likely to overcome the input obstacle. We already have "goggles" and small "flip down" transparent LCD screens that fit directly over the eye that may provide the necessary resolution for activities such as mobile Web browsing at the level of sophistication we have come to expect from our desktop computers and our home entertainment centers.

For the past 30 years, satellite communication has made near-zero cost connections available on a worldwide basis. They work especially well for data streams, but have an inherent delay when used for real-time voice or video communication. Until we find a way to change the physics involved, that is a limitation we are likely to have to deal with.

As is the case for roads, water, and sewer systems, we are likely to be tied to older network infrastructures for some time to come. The physical connections required for many services are expensive, with long terms for return on investment. Until the marketplace stops buying older technologies, we will still find them as part of our communication network. Wireless is not a panacea either. Electromagnetic and optical spectra, although theoretically unlimited, are limited by reality and practicality. In addition, the demand for bandwidth continues to grow at a rate that we have been unable to meet. There are environmental issues as well. We use microwaves to cook food for a reason—that part of the electromagnetic spectrum can be dangerous to the health of living things.

In the near term, expect e-mail to become more and more a relatively easy interface for the transmission of files and media streams of all sorts at the same levels of reliability and speed available today. In the instant messaging arena, expect more and more chat rooms to be netcam compatible—more videoconferencing in addition to traditional text-based instant messaging.

GLOSSARY

Asynchronous Digital communication for which there is no timing requirement. The transmitter signals the start of each.

Attachment A file associated with an e-mail message that is transmitted along with the main e-mail text.

BITNET "Because It's Time Network" began when two universities began using a leased telephone circuit for communications between accounts on their mainframe computers, eventually becoming a worldwide network. European counterpart is EARN (European Academic and Research Network) in 1982.

CDMA (code-division multiple access) Digital, cellular communication technology that uses spread-spectrum modulation techniques. Every *channel* uses the full available spectrum.

E-mail Electronically transmitted messages.

IMAP Internet message access protocol, a protocol for retrieving e-mail messages, developed in 1986 at Stanford University.

IRC Internet relay chat, a system developed by Jarkko Oikarinen in Finland in the late 1980s. IRC allows multiple participants to join in live discussions.

Instant messaging The generic term for any type of messaging system (e.g. IRC) that allows virtually instantaneous synchronous communication.

LISTSERV A commercial product marketed by L-Soft International that automatically distributes messages to a pre-determined mailing list. The term is often incorrectly used to refer to any mailing list server, including Majordomo, a freeware application.

Mime Multipurpose Internet mail extensions, a specification created in 1992 by the Internet Engineering Task Force for formatting non-ASCII character sets and non-ASCII and messages, such as graphics, audio, and video so that they can be sent over the Internet. A variation, S/MIME, supports encrypted messages.

PLATO Programmed Logic for Automated Teaching Operations. A computer assisted instruction system originally developed at the University of Illinois in the 1960s using hardware provided by Control Data Corporation.

POP (post office protocol) A protocol used to retrieve e-mail from a mail server.

SMTP (simple mail transfer protocol) A protocol for sending e-mail messages between servers.

SMS (short message service) An instant messaging system available as a feature on some digital cellular telephones.

Synchronous Communication that uses a common timing signal, that dictates when individual bits can be transmitted. In the context of this chapter, the term is equivalent to "near real time two-way communication."

TALK Command A UNIX operating system command that allows users to chat with each other.

TDD/TTY (telecommunication device for the deaf/teletypewriter) A device using a Baudot (five bit) character set that allows plain text messages to be exchanged using a low bandwidth modem over standard telephone lines. Originally developed for the deaf community.

WAP (wreless application protocol) Originally developed by Unwired Planet, Motorola, Nokia, and Ericsson to develop a standard for wireless content delivery.

CROSS REFERENCES

See *Digital Communication; Internet Architecture; Local Area Networks; Passwords; Public Networks; Standards and Protocols in Data Communications; TCP/IP Suite; Wide Area and Metropolitan Area Networks; Wireless Application Protocol (WAP).*

REFERENCES

AOL signs messaging allies (2002). Retrieved November 2, 2002, from http://www.usatoday.com/life/cyber/tech/ctf788.ht

Barnako, F. (2002). MSN Messenger gaining on AOL. Retrieved August 2, 2002 from http://cbsmarketwatch.com

BlackBerry (2002). Retrieved November 1, 2002, from http://www.blackberry.net

Bucy, E. (2000). *Living in the information age: A new media reader.* Belmont, CA: Wadsworth Thomson Learning.

CREN History and Future (2002). Retrieved May 17, 2002, from http://www.cren.net/cren/cren-hist-fut.html/

CREN history winding down (2002). Retrieved November 1, 2002, from http://www.cren.net/cren/bitnet1/winding-down.html

Garcia, B. (2001). Wireless Web still counting customers. Retrieved November 1, 2002, from http://www.azstarnet.com/public/startech/archive/040401wire6.html

Grant, A. (2002). *Communication technology update.* Newton, MA: Butterworth–Heinemann.

Grubbs, J. (1988). *Digital communications.* Fort Worth, TX: Master Publishing.

Hafner, K. (2001, December 6). The 30-year path of e-mail. *The New York Times: Technology/circuits.* Retrieved November 1, 2002, from http://www.nytimes.com/2001/12/06/technology/circuits/06EMAI.html

ICQ (2003). Retrieved March 27, 2003, from http://www.icq.com

Is wireless instant messaging the future of communication? (2002). Retrieved November 1, 2002, from http://www.internetnews.com/bus-news/article.php/753901

ISOC Internet Standards Programs (2002). Retrieved May 17, 2002, from http://www.isoc.org/standards/

Internet FAQ Archive (2003). Retrieved March 27, 2003, from http://www.faqs.org/rfcs/

Kruczek, T. (1996, January). MQSERIES: I. What is message queuing? *Technical Support Magazine.* Retrieved March 27, 2003, from http://www.naspa.com/PDF/96/T9601006.pdf

McLuhan, E., & Zingrone, F. (1995). *Essential McLuhan.* New York, NY: BasicBooks.

SIP Center (2002). Retrieved November 1, 2002, from http://www.sipcenter.com/

Pew Internet and American Life (2001). Retrieved November 2, 2002, from http://www.pewinternet.org/

Rheingold, H. (2000). *The virtual community.* Cambridge, MA: MIT Press.

Taylor, E. (2002, September 23). Short and sweet. *The Wall Street Journal,* p. R4.

The living Internet (2002). Retrieved May 17, 2002, from http://www.livinginternet.com/

Trillian (2002). Retrieved November 1, 2002, from http://www.trillian.cc/

Why IRC is important (2002). Retrieved November 1, 2002, from http://www.livinginternet.com/?r/rp.htm

E-marketplaces

Paul R. Prabhaker, *Illinois Institute of Technology*

INTRODUCTION

Gartner forecasts that the total value of goods and services transacted through the Internet, by 2004, will exceed seven trillion dollars, with 40% of transactions flowing through e-marketplaces. There are over 1400 e-marketplaces currently in existence. According to the Boston Consulting Group, total e-marketplace revenues in the U.S. will be around $9 billion by 2005, by 2004 B2B e-commerce productivity gains will amount to 1 to 2% of sales, and by 2010 these gains could increase to 6%, roughly $1 trillion (E-Commerce Times, 2001). The Gartner Group estimate supply-chain marketplaces to grow 25-35% over the next 5 years. They also believe that such marketplaces, if properly implemented, should increase ROI by 40% over a 5-year planning cycle. AMR Research estimates that investments in such e-marketplaces will increase at a compounded annual growth rate of 68% over the next 5 years. They also predict that traditional brick and mortar companies will be the primary drivers behind this investment growth. As business-to-business exchanges dominate today's e-marketplaces, this chapter will predominantly focus on B2B e-marketplaces.

An e-marketplace is an online exchange in which organizations and their communities come together to conduct commerce, access content, and collaborate to improve business performance. An e-marketplace is an Internet platform where users can surmount the barriers between countries, businesses, and computer systems (White Paper, n.d.). Simply put, an e-marketplace enables the law of supply and demand to act fully, increasing market efficiencies significantly. The e-marketplace operator acts as an enabler or, at times, a market maker. According to the Chairman and CEO of IBM, Louis Gerstner, e-marketplaces can (a) help online buyers and sellers find each other, (b) attack the inefficiencies of traditional marketplaces, and (c) play important roles in the e-business economy.

Turban (2002) suggests that all marketplaces can be defined based on a three-dimensional product–process–agent framework (see Figure 1). Using this framework, it is useful to see the coexistence of electronic commerce with traditional commerce and various forms of hybrid commerce.

The emergence of B2B e-marketplaces marks a radical change in the development of buyer–supplier relationships. In just a few short years, industry structures that had stood the test of decades of competition and untold waves of innovation are being completely redefined. Some firms have teamed up with former rivals and small new entrants, while eager third parties have announced bold plans to mediate between buyers and suppliers. In many cases, these crucial decisions were made and announced in little more than a week, driven largely by a vision of what could be and a fear of being late to the party. One thousand four hundred such e-marketplaces have thus far been launched or announced. Although some analysts predict that this number will climb to as high as 10,000, others lament that 1,400 is already 1,000+ too many (Spiegel, 2000). Whatever the final number, it is clear that competition to dominate the e-marketplace sector is heating up, as attrition and consolidation are already starting to occur in the B2B space. In the end, what matters is a firm grasp of the key factors that drive the success and failure of e-marketplaces, success being defined as creating value for e-marketplace members and

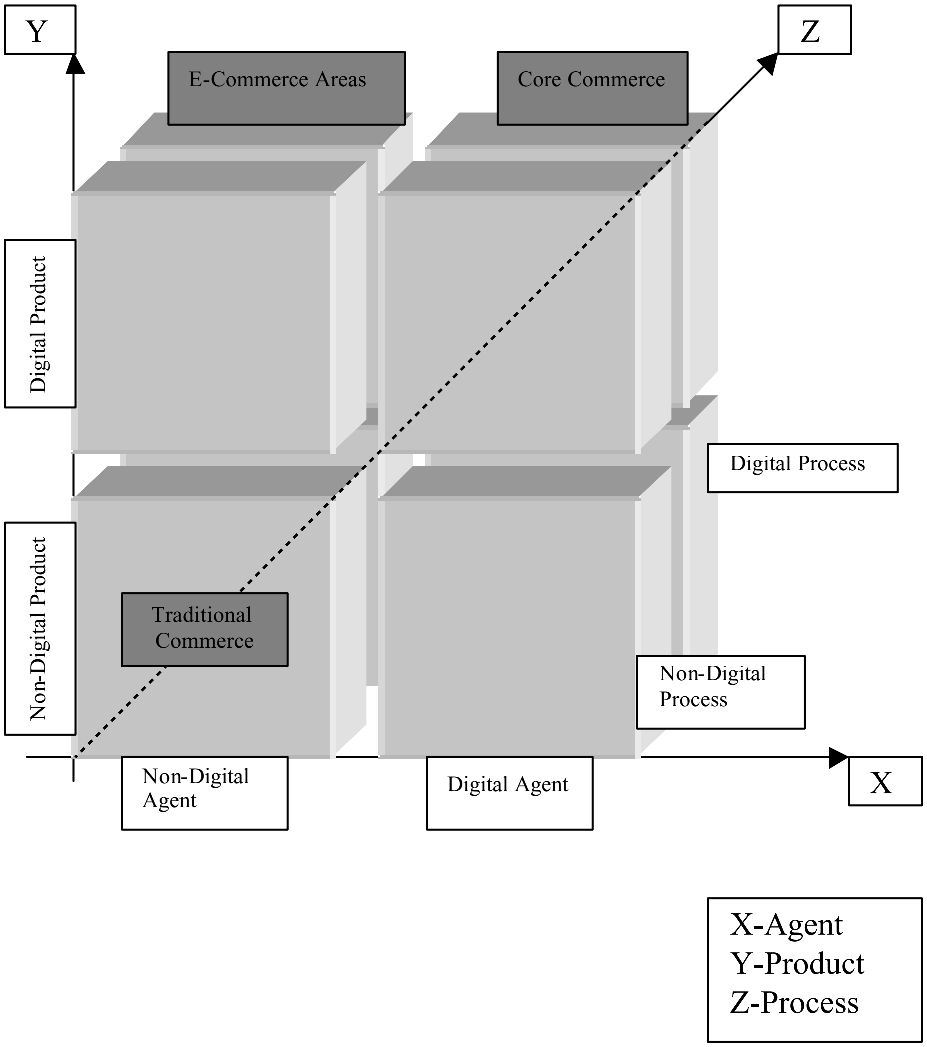


Figure 1: Integrated commerce framework.

their stakeholders on the demand side and the supply side.

The "E" Difference

A marketplace is a physical space where buyers and sellers go in order to match up with each other. This aggregation of buyers and sellers is intended to efficiently match those who have needs for products/services with those who provide the same.

A critical study of the research literature reveals that bricks-and-mortar marketplaces typically need to address five value-drivers if they wish to succeed:

Location

There have to be a sufficient number of interested buyers and sellers located within close proximity of the marketplace. Parties interested in negotiating a market transaction need to physically *go* to the marketplace.

Scale

Transaction *volume* is a key indicator of success. As larger businesses buy or sell more they are able to extract favorable concessions in a traditional marketplace—volume discounts, better credit terms, etc.

Liquidity

For marketplaces to work successfully and profitably, there has to be a defined ratio of buyers to sellers in order to sustain an optimal level of transaction volume and pricing. Physical marketplaces, particularly, have difficulty ensuring this requirement.

Market Rules

The rules regarding price setting, price discovery, negotiations, and transactions should be implemented, taking into consideration the transaction volume generated by a buyer/seller.

Business Relationships

Aside from business transactions, another type of equity is being built up in marketplaces: business relationships. As buyers and sellers repeatedly transact business in a marketplace, they become familiar with each other and before long trust starts building up. That is, transactions are consummated where the terms of the exchange go far beyond

Table 1 Benefits of E-marketplaces

Buyers' Benefits	Suppliers' Benefits
Lower processing costs	Lower costs of customer acquisition
Reduced processing cycle times	Lower costs of sales
Lower product costs through contract buying	Faster time-to-market
Enhanced supplier relationships with collaboration and supply chain management	Greater access to new buyers and markets
Greater access to new suppliers	Benefits of powerful information acquisition capabilities
Improved decision-making	Ability to personalize cost-effectively
Personalized interface	

the marketplace. The "I Win—You Lose" *negotiations* are replaced with a more comfortable and harmonious win–win *understanding*.

One of the key differentiators of an e-marketplace from a bricks-and-mortar marketplace is that the marketplace *comes* to the participants whereas with the latter they *go* to the marketplace. A direct implication of this principle is that three of the above five factors do not apply to digital marketplaces: location, scale, and business relationships. Businesses can buy and sell goods and services through an electronic marketplace regardless of whether they are large or small companies, in near or distant locations, and whether or not they have an existing relationship with the other parties. That leaves us with two critical success factors for running an e-marketplace: liquidity and market rules. A third success factor, exclusive to digital marketplaces, is the ability to connect internal business processes directly to an e-marketplace. Although this is a critical e-marketplace ability, it is also a challenge in the sense that it needs to be addressed proactively. Vertical e-marketplaces particularly leverage this capability to connect directly.

E-marketplaces and traditional marketplaces perform many functions to establish assurance and reliability. As is obviously the case, commerce, "e" or not, is based on trust. Thus, it is crucial for e-marketplaces to embed trust in their platform via concepts such as assurance and information integrity (http://www.informationintergirty.ortg). They should also have carefully thought out strategies for managing risks inherent in e-marketplaces.

Consider the industrial chemical industry. E-marketplaces have gained quick acceptance in this industry, where 65% of business is conducted through auctions on trade exchanges such as ChemConnect.com, e-Chemicals, and CheMatch.com. E-marketplaces cut through the slow and inefficient traditional system of sales visits and trade shows where business is conducted based on long-term relationships between sales reps and purchasers. Forrester expects $128 billion to flow through these chemical e-marketplaces by 2003.

Forrester predicts that between 45 and 75% of B2B e-commerce will migrate to e-marketplaces over the next few years. The largest impact will be in the computing and electronics, shipping and warehousing, and utilities industries, where Forrester predicts that more than 70% of online trade will go through e-marketplaces. In time even markets for customized products will drift to e-marketplaces. Given a sophisticated enough network, even customized products can be auctioned like commodities.

Benefits of E-marketplaces

E-marketplaces are sources of significant benefits for business organizations (see Table 1). They are used to protect market shares, add new channels, move brands online, and serve as a strategic entry point for international markets and are ubiquitously accessible means to communicate with employees (E-Marketplaces, n.d.).

A key benefit of e-marketplaces is the ability to collect buyer information and to leverage this information in the marketplace to secure an advantage. Personalization can be of great value to consumers (Häubl & Trifts, 2000) and can be used to influence them (Häubl & Murray, 2002).

Types of E-marketplace Buyers

E-marketplace users vary on many dimensions: different purposes, different approaches, and different expectations. This concept is not very different from traditional market segmentation. Customers' behavior does vary in several ways, for several reasons. To the extent that firms can understand and leverage those differences in their marketing approaches, they can benefit from marketing efficiencies. Although customers may differ on numerous dimensions—demographics, psychographics, attitudes, loyalty, etc.—it is their market behavior that ultimately matters. E-marketplace buyers can be classified into the following behavioral segments:

Baseline Buyers

These buyers are just getting started in e-marketplace procurement. Their initial objective is to reduce transaction costs associated with maintenance, repair, and operations (MRO) purchases. On average, these buyers spend $5.6 million on transaction fees, integration software, and internal staffing.

Spot Market Dabblers

These buyers use e-marketplaces to make spot purchases of direct materials to help manage inventory and avoid shortfalls. On average, these buyers spend $10.7 million on new software installation and related consultant fees.

Aggressive Spenders

These buyers tend to use the Internet and other networks to manage all their purchasing. On average, these buyers spend $22.9 million mainly on consultant fees for implementing their e-marketplace approach.

GENESIS OF E-MARKETPLACES

According to Forrester Research, over the next 5 years, business purchasers will spend from $5.4 to $22.9 billion each to integrate into business-to-business e-marketplaces. Similarly, Jupiter predicted that businesses would increase their spending on business-to-business e-marketplaces from $2.6 billion in 2000 to $137.2 billion in 2005. According to a Boston Consulting Group study, by 2004, B2B e-commerce will bring about productivity gains equivalent to 1% to 2% of sales. By 2010, that figure could grow to 6%.

Early Stages

Early e-marketplace development was characterized by attempts to set up exchange marketplaces, charging a transaction fee for matching buyers and sellers (Berryman & Heck, 2001). More than 1,000 such e-marketplaces were created for several industries. Most of the players in this phase were new entrants that defined a new way doing business. Unfortunately, most of these marketplaces failed to understand the lifeblood of such a marketplace: liquidity. In other words, a few large enterprises generate most of the volume necessary. The rest will struggle. Hence, the transaction-fee-based marketplaces have languished.

The next developmental stage of e-marketplaces had a different business model. Here, enterprises banded together into consortia with their trading partners and even competitors. Most of the players in this phase were brick-and-mortar companies that leverage from ideas in the first phase. A well-known example of this is the GM/Ford/Chrysler joint venture called Covisint. Since then over 100 similar ventures have been started in different industries. The value driver for these second-wave marketplaces was to reduce bid/ask spreads and to bring down transaction costs by matching buyers with suppliers and enabling suppliers to trade with one another to streamline and make efficient the supply-side marketplace. Unfortunately, the consortia, generally speaking, have not been very successful. Other than auctions, most other "products" in such consortia are not profitable.

E-sourcing

Sourcing provides the single greatest opportunity to reduce costs, streamline processes, and enhance overall responsiveness. It is one of the earliest stages in a supply chain. There are three reasons that e-sourcing is a crucial part of any supply chain. First, purchased products and services are the single largest expense in most organizations, accounting for 50 to 55 cents of every dollar earned in revenue. Second, reductions in procurement costs translate into a dollar-for-dollar increase in profits. By contrast, external factors such as overhead, cost of sale, and profit margins dilute improvements in other functional areas such as sales. Third, sourcing has a multiplicative effect on total costs. A 2% cost reduction in the initial sourcing cycle can yield a 14% reduction in the end cost of a new product or service

Effective sourcing has three dimensions along which investments in sourcing can pay off:

Streamlining and automating processes for the acquisition and management of *indirect* items such as office equipment and computer supplies.

Developing, monitoring, and executing supply-chain and logistics processes for management of *direct* production materials.

Designing, implementing, and nurturing the *strategic* aspects of sourcing.

70 to 80% of the total cost, quality, and structure of products are determined by the end of the sourcing and design cycles. Research across industries shows that 75% of companies believe that their ability to control the cost or quality of a product is practically nonexistent after the initial sourcing process. In the high-tech industry, about 80% of product cost is determined from and attributable to decisions made on the product designer's table. So the information to evaluate strategic sourcing options (or pursue alternative design considerations) is available well before actual sourcing is initiated.

There are different aspects to e-sourcing. Organizations can dramatically improve the processes, cycles, and results of sourcing engagements by leveraging Internet-based technologies to automate and manage the sourcing process. However, sourcing is a complex process that technology alone cannot address. Effective sourcing solutions are those that can integrate advanced negotiation technologies with sourcing methodologies and product-category intelligence and reach upstream into product design stages.

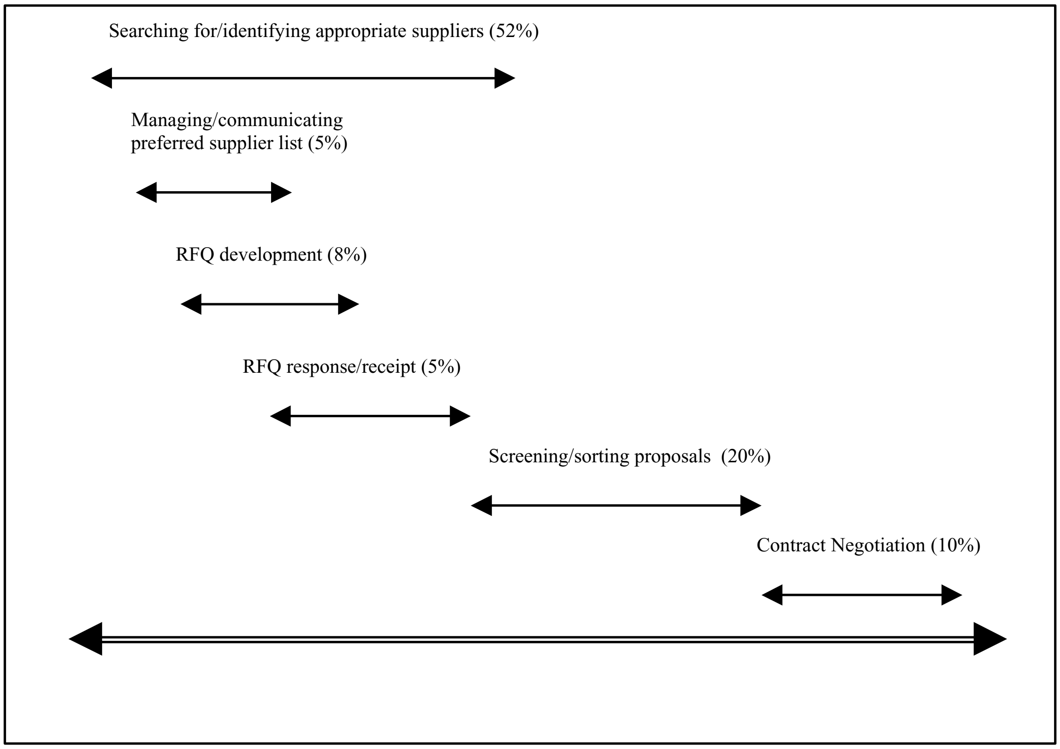
In traditional sourcing, more than half the sourcing cycle at a typical company is currently dedicated to identifying and vetting suppliers (see Figure 2). Another 20% of the total sourcing cycle involves the screening, sorting, and reviewing of supplier proposals. Traditional sourcing approaches are fundamentally electronic data interchanges (EDI) intended to automate a variety of related documentation. Unfortunately, EDI has high set-up and operational costs and burdens that make it ill affordable to a majority of players and lacks the functionality for online negotiation.

There are several benefits of e-sourcing:

Identifying, qualifying, and negotiating with an increased number of suppliers, creating more competitive bidding environments;

Negotiating an average 5% to 20% unit price reduction;

Shortening sourcing cycles by an average of 25% to 30%;


Searching for/identifying appropriate suppliers (52%)
Managing/communicating preferred supplier list (5%)
RFQ development (8%)
RFQ response/receipt (5%)
Screening/sorting proposals (20%)
Contract Negotiation (10%)

Average sourcing cycle = 3.3 months to 4.2 months

Figure 2: Typical sourcing cycle.

Reducing time-to-market cycles by 10% to 15%;

Lowering process costs for sourcing engagements;

Improving quality levels for the goods and services being sourced;

Increasing access to technology and service innovations through improved collaboration;

Applying strategic sourcing to a broader range of products and services; and

Promoting knowledge-sharing and standardization of sourcing best practices across the enterprise.

E-procurement

E-marketplaces have their roots in e-procurement and evolved as a consistent response to the myopic "buy-side/sell-side" mentality (see Figure 3). The latter approach emphasized the efficiency of each side, sometimes at the cost of inefficiencies on the other side. E-marketplaces revolutionize the exchange concept of a market. A marketplace is not just a place to execute an exchange transaction between two parties; it is a place where multiple parties with an integrated set of goals follow commonly agreed-to procedures, where the exchange transaction itself is an incidental matter. The singular focus on transactions by early e-marketplaces brought unsustainable success to them. Later entrants leveraged from the lessons that emerged. The key here is the integrated, collaborative platform that the marketplace provides. E-marketplaces, therefore, provide their value-add in the platform of integrated market operations. The type of collaboration encouraged in a properly run e-marketplace moves away

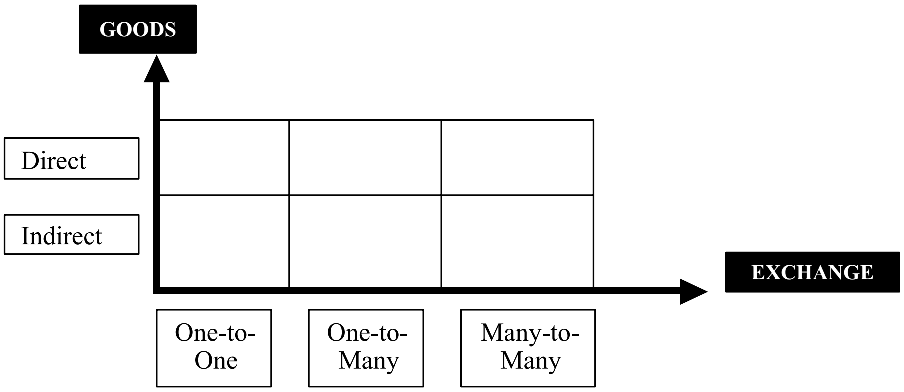


Figure 3: E-procurement to e-marketplaces.

from emphasizing cost and delivery to innovation and reward. In other words, the market partners do not try to eke out more concessions from each other but work together to elevate the success of the entire supply chain (White Paper, n.d.).

E-procurement enabled buyers and sellers to automate their interactions and collaboration. Purchasing processes are linked to back-office enterprise systems. As buyers and sellers integrated their functions, the supply chain became more transparent. According to *Logistics Magazine*, e-procurement can cut distribution costs below the typical 8% of sales. The Aberdeen Group claims to have evidence that processing costs are reduced from $107 to $30 and product costs are reduced by from a minimum of 5% to a maximum of 20%. At the same time, the procurement cycle can be trimmed from typically 7.3 days to 2 days. Ariba, often regarded as the pioneer of e-procurement software, advocates purchase order transaction cost reduction from $150 to $25 for most large U.S. corporations.

Despite these impressive gains, it became increasingly clear that e-procurement is only part of the solution, for two reasons. First, e-procurement applications are more commonly applied to indirect goods—such as office equipment, stationery supplies, and repair services—when such goods account for less than one-third of manufacturing expenses. Only a limited amount of items that a corporation has a need for can be bought through an e-procurement channel. The challenge to e-procurement vendors to extend their solution into the direct materials territory was insurmountable in most cases, because such a shift required domain expertise in the industries served. Second, e-procurement is a one-buyer/one-seller model, whereas most companies do business with multiple vendors and vice versa.

The many-to-many model was quickly accepted in the B2B world, as the benefits were obvious: enabling more buyers and sellers to interact, making it possible for multiple systems to connect, and eliminating the need to implement multiple platforms to link multiple partners.

The many-to-many e-marketplace functions as a central online market where buyers and sellers can share information, initiate transactions, integrate supply-chain goals, and collaborate. The value of an e-marketplace increases as the quantity and quality of its buyers and sellers also increase.

Ultimately, e-marketplaces are not about transactions or collaboration; they are about a platform of *trust*. Raisch (2001) suggests that e-marketplaces will evolve beyond transactional platforms, and even beyond knowledge exchange platforms, to *value trust networks*. Trust between parties, when interlinked, forms a trust network. Such trust networks, it is suggested, could become a critical platform for creating online value. Trust is a complex construct, based on familiarity, integrity, security, and privacy (Prabhaker, 2000).

E-MARKETPLACE CASES

There are several e-marketplace successes, just as there are several e-marketplace failures. Let us look at some e-marketplace cases to understand their uniqueness and patterns for success.

TradeWeb

A description of TradeWeb, paraphrased from their Web site, is as follows:

> TradeWeb is the world's leading online trading network for fixed-income securities. TradeWeb offers a complete spectrum of services. You can view the largest fixed-income markets in real time via TradeWeb's premier Market Data. You get unprecedented depth and breadth of information as well as tight bid and offer indications. TradeWeb's dealer-to-customer platform provides the deepest pool of liquidity for fixed-income products. Today, the network delivers the market-making power of 18 of the world's leading primary dealers to over 1,000 of the largest buy-side institutions. Since inception in 1998, more than $10 trillion in bond trades have been executed over the TradeWeb network. More than $40 billion in securities change hands through TradeWeb every business day. Get live, executable prices from multiple dealers and markets in seconds.

TradeWeb is an online bond exchange. TradeWeb enables a bond buyer to request quotes from multiple brokers simultaneously. Prices pop up in 10 seconds and are updated a million times a day (Weinberg, 2001). The fund manager clicks on the most appealing offer and receives confirmation a second later. Compare this online exchange process to the manual offline process it is replacing: bond buyers had to call several brokers for prices, the brokers called their trading desks who relayed the information to the brokers, and the brokers then passed the price information back to the investors. The offline process often took so long that the prices quoted were not current by the time the investors received them! The icing on the cake here is that the cost of asking and receiving price quotes decreases dramatically thanks to online exchanges.

What makes a TradeWeb type of online e-marketplace so effective?

One key seems to be that successful e-marketplaces limit themselves to the role of an efficient online platform for buyers and sellers to collaborate *without trying to make a market by themselves*. Those e-marketplaces that tried to do the latter by creating disruptive Internet marketplaces generally failed as they discovered that an essential requirement for such disruptive business models to succeed was to garner buy-in from archrivals within an industry. In other words, disruptive technologies need to be funded by top players within an industry. On hindsight, it is easy to see why such archrivals hesitated to feed their trade secrets into a common technology, disruptive or not. As an example, consider Dial-A-Truck Inc., which is in the business of electronically matching rigs with cargo. Rival sites with the same business goal were unsuccessful, as they threatened to eliminate brokers, who for decades had matched trucks with cargo. Dial-A-Truck Services, on the other hand, gives truckers, customers, and brokers a quick and easy way to pair up.

Can e-marketplaces succeed by being industry neutral? Yes, as long as they can find a way to generate liquidity by leveraging their process infrastructure. Thus, the second key seems to be the ability and willingness of e-marketplaces to provide process improvement services. For instance, Toys "R" Us uses amazon.com's infrastructure to sell toys online. Instead of trying to put brick and mortar retailers out of business, amazon.com may be better off using their online process infrastructure to serve as the retailers' Web distributor.

Transora

A single point of connectivity. The leading global e-marketplace for the consumer packaged goods industry, Transora offers an integrated array of collaborative solutions to eliminate inefficiencies throughout the supply chain to deliver breakthrough value.

Transora is an e-marketplace for consumer packaged goods (CPG) companies. It is a successful example of a public marketplace funded by specific partner organizations. In the case of Transora, investors include Campbell Soup Company, Bristol-Myers Squibb, Kraft Foods, Procter & Gamble, Coca-Cola, and Unilever. Funding from partner companies has exceeded $250 million. Such a financial tie-in with an e-marketplace creates a strong incentive to use the e-marketplace often. Transora has used the money raised to make their marketplace more valuable by adding community, content, marketing tools, and supply-chain services.

ChemConnect

ChemConnect is the leading online chemical and plastics global marketplace. Their mission, paraphrased from their Web site, is as follows:

> We know your industry and the challenges you face when buying and selling chemicals and plastics, and we've developed solutions to meet those challenges. Now you can access reliable market information, reduce costs, and increase efficiencies—all of which helps you compete on the worldwide market.

Two giant online exchanges in the chemical industry, ChemConnect and CheMatch, have merged, resulting in a one-stop, $4+ billion chemicals and plastics e-marketplace (Harreld, 2002). The new ChemConnect offers a comprehensive solution including online auctions, a commodity spot and futures exchange, and an electronics communication hub for the automatic transfer of transaction data. ChemConnect has successfully launched a subscription-based revenue model in addition to transaction fees.

FreeMarkets

FreeMarkets' mission, paraphrased from their Web site, is as follows:

> FreeMarkets is the leading global provider of sourcing software and service solutions. Our sourcing software and service solutions help suppliers to win new business and buying organizations to dramatically improve their sourcing process and identify immediate and ongoing savings. Our flexible portfolio of sourcing software and service solutions provide companies with all of the tools they need to dramatically improve their sourcing process and identify fast, measurable savings by conducting strategic sourcing online. Through our sourcing software and service solutions, organizations can identify high-quality, global suppliers, create and distribute detailed Requests for Quotation (RFQ), and structure and execute effective online markets for a wide range of goods and services. The world's leading and largest companies have used FreeMarkets to source more than $35 billion in goods and services, and identify savings of over $7 billion to date

The revenue model for FreeMarkets is simple: 1% to 3% of the value of what clients source. Ultimately, FreeMarkets lets customers decide how to use its technology, either in public auctions or in private auctions. Then it implants itself deep into a client's infrastructure to make it work. FreeMarkets charges buyers based on volume. FreeMarkets' revenue model is based primarily on auction value and consulting services. The auctions are private or public for new, used, or surplus goods. The consulting services are for identifying the buying or auctioning needs of clients. Businesses will want to study the pricing models carefully: what works for one industry may not work for another.

Covisint

> Covisint is the central hub where OEMs and Suppliers of all sizes come together to do business in a single business environment using the same tools and user interface. Covisint's online tools will enable your company to compress planning cycles and enhance supply chain planning. In doing this, it allows you to directly increase efficiency and asset utilization, while ultimately realizing greater profits and shareholder valuations.

Covisint is one of the most famous examples of a public e-marketplace. The intention is to create a supply chain management arena for the entire auto industry, worldwide. Their trading partners include

ArvinMeritor
DaimlerChrysler
Delphi
Faurecia
Ford Supplier Portal
Freudenberg
General Motors
JCI
Nissan

Peugeot

Renault ("Covisint will save (Renault) $280 per car in reduced inventories and faster lead times"—Renault CEO Louis Schweitzer)

Siemens Automotive

Tower Automotive

Yazaki NA

The ambitions, however, may or may not be realized. Different automakers, such as Volkswagen and BMW, are choosing to create private e-marketplaces. Covisint has been a classic case study of the economic benefits of a large-scale public marketplace vs. the downside of consortium limitations of such a marketplace. Thus, a large industrial marketplace may not be assured of success over a company's customized private marketplace.

E-marketplace Case Analysis: PetrochemNext.com

PetrochemNext.com is an online gateway to the world of plastics. The exchange is designed to answer the commerce, content, and community needs of manufacturers, consumers, end-users, distributors, dealers, and others—in short, people whose world revolves around plastics. The portal enables users to buy and sell plastics online in the real-time and dynamic marketplace, access updated technical and commercial information about plastics, and interact with members of the plastics community.

Corporate Profile

Internet ExchangeNext.com Ltd. and the Indian plastics community collectively own PetrochemNext with a strong partnership base and equity participation by leading manufacturers, distributors, dealers, and processors.

PetrochemNext has partnered with five leading manufacturers for an ample supply of polyethylene, polypropylene, PVC, polystyrene, polyamides, ABS, and thermoplastic polyester through the online marketplace.

Internet ExchangeNext.com Limited (EXCHANGE NEXT) is established to help e-enable the businesses in the Indian manufacturing and services sectors. EXCHANGENEXT is currently setting up infrastructure for end-to-end, e-commerce solutions in the business-to-business arena. The company also hosts and supports other Indian e-marketplaces such as papernext.com, steelnext.com, brokerfirstoffice.com, and brokernext office.com.

The exchange was developed by Internextexchange.com with hardware and e-commerce support from IBM, Web hosting by Reliance Infocom, online payment gateway and net banking with HDFC Bank, and customized software for order management system from Tally.

Activities:

PetrochemNext follows an offline accreditation process to select its trading members, who can trade on multiple trading floors in the online marketplace. The trading floors offered currently are as follows:

Offer To Sell In the Offer to Sell section of the marketplace, sellers can create sale offers of specific products at fixed prices with well-defined terms and conditions targeted at the trading community. The products can be catalog products if they are already listed in the marketplace catalog or noncatalog products if not listed. The seller organization can benefit from the combined purchasing power of multiple buyers across the country, receive direct orders placed through this mechanism, and ensure that complete and accurate data are available to the buyers, thereby maximizing sales.

Buy In the Buy section of the marketplace, buyers can easily and quickly locate sellers by searching through offers, communicate with sellers, make informed purchase decisions, and create purchase orders. This electronic purchasing solution is the fastest and most effective way of doing business online, saving time and money.

Auctions Sellers can create Auctions in the marketplace by specifying the type of auction, the minimum auction quantity, reserve price for the product, if any, and other bid control rules. The buyers can browse through the auctions listed, bid in the auctions, and avail themselves of the best bargains in the market.

Open Cry Auctions This type of auction is similar to the public meeting model. All bids under an open cry auction are available for everyone to see. These auctions work well in situations where the prospective buyers are able to bid in the auction without traversing a geographical distance and want to submit counter bids quickly. On PetrochemNext, however, open cry auctions can be conducted for an extended period of time, giving buyers more time to react and submit bids. Anonymity is maintained throughout the auction.

Reverse Auctions Buyers can create Reverse Auctions in the marketplace by specifying the type of auction, the minimum auction quantity, the opening bid price and starting price for the product, if any, and other bid control rules. The buyers can browse through the auctions listed, bid in the auctions, and avail themselves of the best bargains in the market.

Reverse Open Cry Auction This type of auction is similar to the public meeting model. All bids under an open cry auction are available for everyone to see. These auctions work well in situations where the prospective sellers are able to bid in the auction without traversing a geographical distance and want to submit counter bids quickly. On PetrochemNext.com, however, open cry auctions can be conducted for an extended period of time, giving sellers more time to react and submit bids. Anonymity is maintained throughout the auction.

Revenue Model

Transaction Fees. The buyer organization that requests an e-procurement auction is charged a transaction fee of 1% of auction value or Rs.10,000 per auction [about $200], whichever is higher. This covers hosting the e-procurement auction on the buyer's behalf, training his or her invited suppliers for participation in each reverse auction event, providing him or her with a view of the live auction status, and finally sending him or her an analysis at the end of the auction.

The participating suppliers are not charged a fee.

Advertisement. PetrochemNext.com also accepts advertisements on the Web site. Advertisers can currently display time-based advertisements of their products and/or services on PetrochemNext.com in the following sections:

Discussion Forum
Technical Support
Yellow Pages
Classifieds
My Forum
Product Guide
News and Events

CHARACTERISTICS OF SUCCESSFUL E-MARKETPLACES

Markets are about power; the most powerful customers set the terms of trade. Powerful buyers can moderate the pace of the market and drive down prices. The marketplaces can facilitate this. Some marketplace make collaboration easier and allow suppliers to demonstrate their value-added. E-marketplaces can support both competitive and collaborative activities. E-marketplaces also increase market visibility, making poor performance visible—transparency.

Given the above, there are specific rules that need to be considered in identifying characteristics of successful e-marketplaces. First, markets will only support one to three e-marketplaces within any given industry segment. Second, e-marketplaces will need to bridge the gap between falling transaction fees and rising demand for collaborative services. Third, unfortunately, collaborative services, such as supply-chain forecasting and demand planning tools, are difficult and expensive to implement. Fourth, sellers particularly will be looking for ways to ensure that e-marketplaces do not commoditize their products.

Careful examination of successful e-marketplaces reveals the following common characteristics:

1. Liquidity: Large aggregation of buyers and suppliers with multiple static and dynamic pricing mechanisms
2. Transparency
3. Customized transactions
4. Collaboration
5. End-to-end functionality: Catalog, price discovery (auction, fixed, quote, negotiated), requisition, ordering, status, shipment, insurance, logistics, tax computation, bill presentment, bill payment, report, and reconciliation
6. Open technology standards: For implementation of the marketplace itself; especially standard technology and open interfaces for integration with systems of other parties (buyers, suppliers, banks, logistics, escrow, catalog/content providers, etc).

All six characteristics have a direct bearing on market efficiencies. For example, greater price transparency will lead to a lower price via increased efficiencies in processing price information. However, more efficient integration between buyers and sellers will require collaboration and open technology standards. Similarly, customized transactions provide a value of their own.

So what sets the successful e-marketplaces apart from those that are not? The extent to which they leverage the above characteristics. These characteristics in reality define the market space for e-marketplaces (see Figure 4).

E-marketplaces have been largely buyer-incentivized and buyer-driven. In that sense they polarize the model of a marketplace. The value proposition of e-marketplaces to suppliers has so far been weak. True e-marketplaces should go beyond merely squeezing suppliers' prices to reducing costs and assets for all participants in the supply chain. Successful e-marketplaces need to embrace a business model that calls for going beyond simplistic transaction-only trading networks to providing a collaborative environment linking multiple trading networks, individual buyers, suppliers, and service providers.

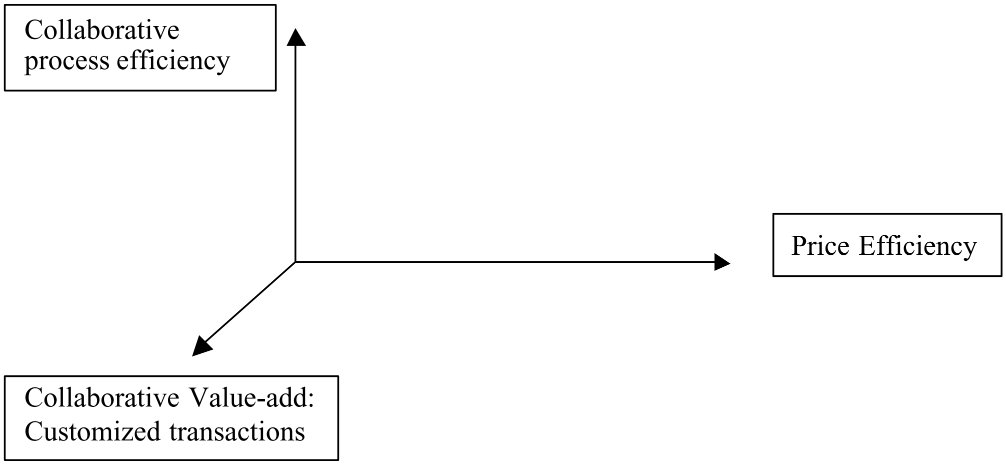

Collaborative process efficiency
Price Efficiency
Collaborative Value-add: Customized transactions

Figure 4: E-marketplace space.

Successful e-marketplaces generally exhibit a key capability to deal with known e-marketplace pitfalls, such as the following:

E-marketplaces that fail to deliver a balance of benefits to suppliers and buyers will find it difficult to generate the threshold liquidity to survive.

E-marketplaces that do not contribute to an enhanced relationship between business partners will not deliver enduring value.

E-marketplace implementations that are based on closed or proprietary architecture limit a company's ability to link with other participants, thus reducing the overall capabilities it can provide and value it can deliver.

E-marketplace fraud is a key issue. Auction fraud accounts for nearly 43% of all online fraud. Nondeliverable merchandise and nonpayment accounted for 20.3% of complaints and Nigerian letter fraud made up nearly 15.5% of complaints. More than 49,000 complaints were filed with the Internet Fraud Complaint Center in 2001.

E-marketplace Guiding Principles

Based on research on the prevailing business models, ownership structures, marketing tactics, service offerings, and strategic positions in the e-marketplace arena, we can define a set of guiding principles to help e-marketplaces and their participants develop successful B2B strategies. First, e-marketplaces are all about industry transformation and collaboration across the value chain and extended enterprise. Second, as a follow-up, lower product prices get attention but are not where the real value lies. Third, e-marketplaces must deliver value to the core businesses of member organizations. Without delivering core value to all participants (win-win for all), the e-marketplace will ultimately lose. Fourth, e-marketplaces replace inventory with information. Integration and technology "plumbing" will help deliver the value promised by e-marketplaces. Without appropriate technology foundations, e-marketplaces will never be able to replace inventory with information. Fifth, the capabilities required to operate and use e-marketplaces require implementation and integration of many technology components across a variety of business processes and functions. This means multiple technology vendors integrated with people and processes. Finally, each company will need to leverage multiple e-marketplaces of different types and will need to participate in other forms of B2B. This will allow companies to maximize benefits across the goods they purchase, processes, and geographies.

Based on the above principles, there are five critical principles that e-marketplaces and participant companies can use to help guide their strategies and tactics:

Operating Structure and Ownership Long-term success will require that e-marketplaces attract and retain a critical mass of transactions. This will require that the e-marketplace offer value and a level playing field to all participants. Third-party e-marketplaces offer participants a neutral trade environment. However, in industries where power is concentrated in the hands of large brick-and-mortar organizations, brick-and-mortar-led consortia may have the market-making power to reach critical mass and beat neutral third parties. For brick-and-mortar consortia to succeed, new mechanisms will be needed to guarantee a fair environment to all participants.

Governance Long-term success will hinge on how well governance structures can guide and regulate e-marketplace management—balancing the interests of shareholders, members, and outside interests (e.g., the U.S. Federal Trade Commission [FTC]).

Scale The e-marketplace will only succeed if it can maintain transaction volumes that satisfy scale considerations for buyers, suppliers, and intermediaries alike. It must provide buyers with leverage, geographic coverage, and, in some cases, a broad array of available products. It must provide sellers with sufficient economic scale to offset the resources invested and intermediaries with sufficient reason to play.

Regulatory Compliance The e-marketplace must walk the fine line of maximizing benefits and leverage to its members and owners without attracting FTC scrutiny that would hobble implementation efforts.

Technology Enabling capabilities beyond simple transaction matching requires that the buyers, the sellers, and the e-marketplace itself build robust, interconnected systems that span a number of processes, systems, enterprises, and industries. E-marketplaces that support deep integration with their members and can help their members implement and integrate technology will have a competitive advantage, and will become "sticky."

These five principles can help companies looking for guidance in e-marketplace selection separate the eventual winners from the inevitable losers. It is important to note, however, that each e-marketplace is a fragile ecosystem—made up of buyers, sellers, and various intermediaries. As such, the success of each e-marketplace is dependent on the success of each of its participants.

TAXONOMY OF E-MARKETPLACES

E-marketplaces may be classified based on their market-sector coverage, vertical or horizontal. They can also be classified by who has access to their services, private parties or the public. Thus, vertical e-marketplaces may be public or private, and so may horizontal ones.

Vertical E-marketplace

Vertical e-marketplaces strive to provide a particular market sector with a network platform for an automated, Internet-based transaction model for made-to-order products.

Advantages for participants include

Better supply-chain management,

Online strategic sourcing,

Reduced costs,

Enhanced vendor relationships, and

Improved inventory management of raw and finished goods.

Examples

Covisint (Automobile industry)
WorldWide Retail Exchange (WWRE) (Retail industry)
GlobalNet (GNX) (Retail industry)
Transora (Consumer Packaged Goods [CPG] industry)

Horizontal E-marketplace

Horizontal e-marketplaces strive to provide customers with a generic platform for pan-industry products and value-added services.

There are several products and services that are essential for the efficient production and delivery of goods across industries. Examples of such pan-industry products would range from stationery products to logistics services, payment services, etc. Example: Ariba for MRO products.

Private E-marketplace

Private e-marketplaces strive to provide benefits for specific supply-chains/consortia/enterprises.

These e-marketplaces include e-procurement and e-distribution modules. They enable the consortia/enterprises to deliver the right products in the right quantity to the right customer at the right time. To that extent private marketplaces need to be treated as separate e-businesses by the participants: that is, the participants need to invest, nurture, and grow these as though they were in-house business units.

Public E-marketplace

Public e-marketplaces strive to provide benefits for any and all interested enterprises/supply-chains/consortia.

Public e-marketplaces are stand-alone marketplaces that are distinct profit-making entities. Their business models revolve around their ability to leverage online technologies in delivering value to buyers and sellers in their exchange transactions, processes, and collaborations.

Another way to classify e-marketplaces is to look at their functions:

Procurement Marketplaces

Aggregate suppliers and allow customers to search/compare offerings of different vendors (e.g., bCentral's Buy & Sell page).

RFQ/RFP Matchmaking

Here, buyers submit their need specifications through a request for quote (RFQ) or request for proposal (RFP); interested vendors then submit a quote/proposal for the work.

Auctions

These are demand-based exchanges allowing buyers to bid against each other for items offered. The final selling price is usually determined by the highest bid received for the item.

Barter Exchanges

These are e-marketplaces where businesses trade their products and services. No money exchanges hands.

Trading Networks

Trading networks, or trade exchanges, are a basic type of e-marketplace focused on reducing purchasing costs. A unique characteristic of trade exchanges is the bid/ask pricing mechanism. They are designed to squeeze procurement benefits from *processes* and *trading partners*.

Hybrid E-marketplaces

Hybrid e-marketplaces are an integrated, customized solution for enterprises that need the efficiency of e-marketplaces and would also like an enhanced brand-building customer experience.

Benefits of Hybrid E-marketplaces

Protect premium brand: Achieved by controlling what products are sold through private and public exchanges and dictating the business rules for each type of transaction

Increase market share: The ability to differentiate between public e-marketplace customers and private exchange customers enables effective market segmentation. A carefully implemented segmentation strategy will usually result in an increase in market share.

Increase revenue: Reaching more customers with more personalized offerings results in an increase in revenue and profits.

Reduce system costs: A single integrated interface for channel management and personalization reduces administrative costs

VALUE-ADD DIMENSIONS OF E-MARKETPLACES

E-marketplaces generally add value along multiple dimensions. First, they leverage network effects. Essentially, more connections mean more value, and better connections mean more value. The value creation domain here is based on inter-firm relationships. Consumer information that was maintained by companies with their own computers now resides in giant databanks accessed by globally networked computers (Kakalik & Wright, 1996). Second, e-marketplaces create and deliver value-add services. Successful e-marketplaces present more than transaction services; they also provide business services that complement transactions. For instance, firms combine data from their warranty-card databases with credit-bureau files to dramatically improve their insights into consumers' lifestyles (Scott & Shermach, 1998). Services such as credit services, payment services, taxation, shipping, and documentation would be such value-add services. More importantly, e-marketplaces consolidate all these business services into one site so buyers and sellers can use a single site for transactions and *all* business services they require. Third, they provide a collaborative infrastructure capability. As e-marketplaces evolve, many back-office applications migrate to the Web site. Organizations transfer responsibility for processes, such as business intelligence, product planning, and promotional activities, to the e-marketplace. By moving this workload to the e-marketplace organizations are better able to collaborate with their sellers and buyers, jointly managing their shared supply chains across multiple platforms without the need to implement additional applications. Finally, time saving and reduction of inventory-carrying costs are the greatest benefits of digital marketplaces. Online auctions, a form of e-marketplace, have cut down procurement time from about 2 months to a matter of hours for Sprint. Process efficiencies are improved both outbound and inbound. In its first reverse auction, Sprint received 250 bids in the very first hour! Such gains, however, come only for those prepared to capitalize on them. For instance, companies need to be able to link data generated online with back-office systems such as purchasing, inventory, and accounting. Otherwise, information must be updated and entered again, eliminating any gains form digital marketplaces.

E-marketplaces Add Value to Supply Chains Along Three Dimensions

Clearly, improved collaborative communications is a key, as it results in better market efficiencies. What is Web-based collaboration? It is more than just managing inventories and maintaining communications; it is a Web-centric relationship management system characterized by multiple network computing systems and communities of people. See Figure 5.

E-marketplace Benefits

A different perspective on e-marketplace value-add is seen if we look at different *levels* of value rather than different *recipients* of value. Electronic marketplaces result in certain direct value-adds that are denoted first-order. If managed properly, these value-adds should lead to measurable results that are second-order benefits. Beyond this, if the value-adds are combined with other capabilities they can result in competitive advantages, which are called third-order benefits.

First-order benefits:

Improved supply-chain visibility,
Reduced inventory costs,
Reduced procurement costs,
Improved order and approval processes,
Reduced transaction costs, and
Relationship management [or perhaps this should be in second order].

Second-order benefits:

Improved asset utilization,
Better capacity planning and utilization,
Higher labor productivity,
Increased economies of scale,
Decrease in product development times,
Increased market share,
Higher profitability, and
Reduced time to market.

Third-order benefits:

Higher service levels,
Easier product customization,
Increased customer loyalty, and
Lower marketing costs.

E-marketplaces are realizing that they need to offer more than efficient commerce capabilities to succeed in

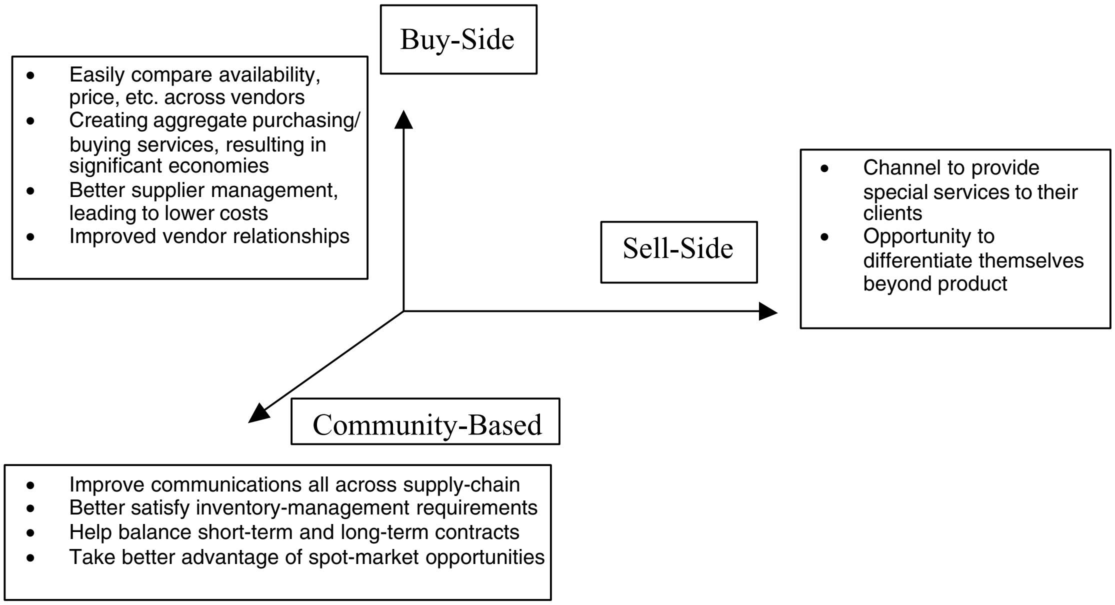

Buy-Side
• Easily compare availability, price, etc. across vendors
• Creating aggregate purchasing/ buying services, resulting in significant economies
• Better supplier management, leading to lower costs
• Improved vendor relationships
Sell-Side
• Channel to provide special services to their clients
• Opportunity to differentiate themselves beyond product
Community-Based
• Improve communications all across supply-chain
• Better satisfy inventory-management requirements
• Help balance short-term and long-term contracts
• Take better advantage of spot-market opportunities

Figure 5: E-marketplace value-add space.

the tough competitive environment: they need to offer genuine value through those capabilities. Additionally, these values have to be realized from the perspective of all participants, not just buyers. Horizontal capabilities include personalization, communities of interest, portal capabilities, and the capability to handle RFQs and RFPs. The value in these capabilities is not just transaction support but also the ability to embed the role of e-marketplaces more deeply in the trading partners' relationships with each other. Vertical capabilities, on the other hand, include add-on services that let the marketplace control the end-to-end trading cycle. Such services focus on settlement, fulfillment, and returns processing. Companies in specific industries, such as banking, financial services, insurance, and shipping, all have a vested interest in getting their piece of e-marketplace trade.

For instance, consider settlement services:

E-marketplaces need credit validation services to verify that buyers have sufficient funds available to back their purchases. Also, in certain cases there is a need to create escrow accounts to hold goods until funds are actually received.

Trading partners need lines of credit for operating capital and to expand their businesses. E-marketplaces could extend lines of credit to buyers in order to settle the buyers' accounts while charging the buyers interest or fees. Timely settlement for a fee!

Consider fulfillment and returns processing services:

Many low-tech trading partners do not have the infrastructures needed to handle these functions. E-marketplaces can take on more responsibility for delivery and returns. Insurance can also play a role here.

Invoicing, bill presentation, and payment are key services for marketplaces. As buyers and sellers use a marketplace for trade, it is logical that they use the same marketplace as a conduit for these services.

Key: the vertical services apply not just to new marketplace entrants but also to brick-and-mortar companies that are adding e-marketplaces to expand their businesses. *The challenge for any marketplace is in integrating all of these services into its offerings.*

CONCLUSION

E-marketplaces are powerful vehicles to achieve a competitive advantage in today's fiercely competitive industry sectors. Unfortunately, many e-marketplaces have missed the point. The focus, incorrectly, is on transactional efficiency, instead of on improving business processes to unlock additional value. E-marketplace-enabled information sharing and collaboration are the keys to this value. Also, effective e-marketplaces recognize that individual managers and individual companies make buy/sell decisions, not entire industries. Thus, attempting to transform the procurement practices of entire industries utilizing e-marketplaces generally is not successful.

Electronic marketplace exchanges are here to stay because they make great business sense (Oliver, 2001). Based on sensible business models, e-marketplaces can add game-breaking capabilities to a firm's competitive arsenal. Both vendor and customer firms stand to benefit from good e-marketplaces. Misused, such e-marketplaces can put the same vendors and customer firms in a deep competitive hole. This makes a strong case for carefully harnessing the power of e-marketplaces.

FUTURE ASPECTS OF E-MARKETPLACES

It is important to bear in mind the foundational aspect of e-marketplaces: those that are able to create a network effect—delivering greater value as increasing numbers of participants join and collaborate online—will outlast others.

Gains that result from increased transactional efficiency alone will not be sufficient to convert buyers and sellers to buy into the concept of e-marketplace. True gains from marketplaces will result not from trading but from better access to information. In other words, collaboration based on sharing of information over the Web seems to be the real value of e-marketplaces. For example, supply and demand forecasts, inventory level reports, and effect of price on demand are some ways to leverage collaboration on line.

Also, based on the same hindsight, it is possible to arrive at some key conclusions. The e-marketplace field is enormous and continually expanding. Many players will appear but many more will disappear, as value-creation and industry transformation become the central focus of e-marketplaces. The first conclusion is that *fewer, larger e-marketplaces will be required to realize significant value.*

The multi-trillion-dollar projections from a variety of industry analysts indicate that B2B is just beginning to hit its stride. E-marketplaces and B2B participants should adopt a strategic, long-term view. E-marketplaces are an outcome of the interest in jointly managed supply chains, resulting in greater service integration and more collaboration. The second conclusion, therefore, is that, increasingly, *competition is between supply chains rather than between corporations*. Guly (1998) suggests that this new economy has new rules. In the traditional economy *value* typically was a result of scarcity, whereas in the new digital economy value comes from plenitude. This important observation points to the foundation for business strategies of the future. Managing supply and demand for products is not as important as managing networked relationships of products, probably owned by different companies.

Operational efficiency is the primary driver for e-marketplace development. Hence, the third conclusions is that *companies must wrap a business process around the initiative rather than adopting it as a technology, or the initiative will not be successful*. One of the most enduring lessons from the rise and fall of dot-com mania is that the rules of business have not changed much at all. Initially, the corporate motivation for participating in e-marketplaces will be to save money. However, as a final conclusion, *corporations should treat their participation in e-marketplaces as a strategic asset and suppliers must prepare themselves for new rules in these dynamic marketplaces.*

APPENDIX: E-MARKETPLACE ASSESSMENT TOOLS

E-marketplace Orientation Index

Should we "e" our marketplace? Although e-marketplaces have significant benefits, not all companies are endowed with the market/organizational environment to reap the full rewards e-marketplaces. Here is a quick and easy way to compute an (index) of e-marketplace orientation:

Score the business on a scale of 1 to 5 on each of the following items:

Target Customer: Who Are the Customers?

E-marketplaces can be designed to cater to business-to-business, business-to-customer, business-to-consumer, or a variety of hybrid models. Consumer-based goods and services tend to be sold directly through the Web, rather than through e-marketplaces. Auction sites are obvious exceptions.

[1 = individual consumers; 2 = consumer groups; 3 = individual customers (resellers); 4 = business customers (resellers); 5 = business customers (value-add resellers)]

Market Coverage: Where Are the Customers?

The more homogenous and local the market, the better brick-and-mortar marketplaces work. The more fragmented and distant the market, the better e-marketplaces work.

[1 = local + homogenous clusters; 2 = local + fragmented market; 3 = regional/national + fragmented market; 4 = distant/international + homogenous clusters; 5 = distant/international + fragmented markets]

Buying Process: How Do the Customers Buy?

The more customers are accustomed to using technologies in their buying process the more natural the fit with e-marketplaces.

[1 = see actual product + buy onsite; 2 = see actual product + buy through telephone/fax; 3 = catalog purchases; 4 = shop offline+ buy online; 5 = shop and buy online]

Market Growth Trend: Is the Market Expected to Grow?

The adaptation to an online marketplace can be stressful on existing business practices and processes. At best, an online marketplace will add, requiring more human and financial resources. This extra resource demand makes sense if the market is in a growth mode.

[1 = declining market; 2 = flat market; 3 = slow growth; 4 = stable growth; 5 = fast growing market]

Organizational Stress: Cost of Adapting to E-marketplaces?

Many other business functions will be affected—not just those that reap the immediate reward of e-marketplaces. Given this situation, it is important to evaluate the cost of the organizational stress generated by participating in e-marketplaces.

[1 = internal stress nearly impossible to deal with; 2 = very difficult in the short term; 3 = can find a way to adapt; 4 = pretty easy; 5 = no stress due to e-marketplaces]

Disintermediation Reactions to E-marketplaces by Vendors/Suppliers?

The impact of e-marketplaces on existing business relationships with vendors/suppliers/customers/business partners can be painful; this would call for an adjustment not just internal to the firm but also by external parties. How these external parties choose to adapt to the realities of e-marketplaces can either make or break participation in such online marketplaces.

[1 = existing vendor relationships cannot be changed; 2 = existing vendor relationships can be changed at great cost; 3 = mandated change; 4 = fairly easy as vendors see long-term benefits; 5 = vendors share in e-marketplace gains]

Interpreting the E-marketplace Orientation Index Score

A score of 1 through 10: E-marketplaces are *not* a natural fit.

A score of 11 through 20: Will realize e-marketplace gains in the long run.

A score of 21 through 30: E-marketplaces are a natural fit. Implement immediately.

E-marketplace Screening Grid

Which e-marketplace is the right one for a business? Address the following five questions to systematically define the ideal e-marketplace for a given business's needs and circumstances:

1. What Types of Businesses Participate in the E-marketplace?

Check out the neighborhood. What companies participate? What types of items do they sell? How successfully? How many buyers? First-time/repeat customers?

2. How Do Customers Buy?

Some e-marketplaces just connect buyers and sellers; others offer all sorts of support services and manage the entire exchange process. What do the customers in this exchange expect?

3. What Will It Cost?

Fee structures and add-on charges for e-marketplace participation vary substantially. What is the breakdown of costs?

4. How Are the E-marketplace Vendors and Customers Evaluated in a Particular Exchange?

What vendor screening criteria does the site use? What customer credit risk evaluation criteria does the site use? Are these criteria consistent with the business philosophy?

5. Who Controls the Presentation of Products on the Exchange?

e-marketplaces have their own policies and standards on product presentations. Ideally, the company should retain maximum control over presentation of its products.

E-marketplace Evaluative Framework

How do we estimate the value-add of an e-marketplace to our business? Score the e-marketplace on the following criteria (Microsoft bCentral) to evaluate an exchange's capabilities and offerings.

1. Easy Product Updates

E-marketplaces should make it easy to update product catalogs, implement special promotions, discounts, etc.

2. Back-End Integration

E-marketplaces that can be integrated into back-office functions (accounting, inventory, etc.) can reduce costs and improve efficiency.

3. Reports

Versatility in reporting options provides a significant advantage in forming marketing and sales strategies. Such reports allow useful information—e.g., location of best customers, buying habits of different customers, identifying product bestsellers—to be gathered and sent to the right person at the right time.

4. Transaction Support:

What level of support does the exchange offer? Credit card payment, shipping options? Customers are increasingly expecting such transaction support services.

GLOSSARY

Brick-and-mortar marketplace A traditional marketplace where the buyers and sellers meet to do business.

B2B (business-to-business) Selling products or providing services to other businesses.

Consortium An association or a combination of businesses, financial institutions, or investors for the purpose of engaging in a joint venture.

E-marketplace An on-line exchange in which organizations and their communities come together to conduct commerce, access content, and collaborate to improve business performance.

E-procurement Procurement through fully automated, Internet-based self service systems that streamline the transactional purchasing process between the buying organization and its suppliers.

E-sourcing Strategic sourcing using Internet tools.

EDI (electronic data interchange) The transfer of data between different entities using electronic networks, such as the Internet. This transfer can occur both within and between organizational entities.

Fragmentation A market condition characterized by many uncoordinated buyers or suppliers. Fragmented markets are inefficient and costly to reach. Fragmentation also exists within large purchasing organizations when procurements are made in smaller, uncoordinated batches, resulting in higher prices.

OEM (original equipment manufacturer) A company that purchases products or other complex components, often for specific applications, from vendors and then adds original technology, hardware or software, and sells the aggregated products/systems.

RFQ (request for quotation) An invitation for suppliers to bid on clearly specified products or services. Exchange technology can combine multiple RFQs to produce lower bids and increase sales efficiency.

ROI (return on investment) The income that an investment provides in a year.

Strategic sourcing A disciplined approach used to optimize sourcing decisions for goods and services.

Supply chain The total sequence of business processes, within single or multiple enterprise environments, which enable customer demand for a product or service to be satisfied.

CROSS REFERENCES

See *Business-to-Business (B2B) Electronic Commerce; Consumer-Oriented Electronic Commerce; Click-and-Brick Electronic Commerce; Electronic Commerce and Electronic Business; Electronic Data Interchange (EDI); Electronic Procurement; Supply Chain Management*.

REFERENCES

Berryman, K., & Heck, S. (2001). *E-marketplaces*. New York: McKensey and Co.

E-Marketplaces (n.d.). Retrieved April 3, 2003, from http://www.research.ibm.com/irl/projects/market

Guly, C. (1998, October 26). OECD summit tackles e-com. *Computing Canada, 24*(40), 1,4.

Harreld, H. (2002, January 10). B-to-B titans ChemConnect, CheMatch to merge. *InfoWorld*.

Häubl, G., and Murray, K. B. (2003). Preference construction and persistence in digital marketplaces: The role of electronic recommendation agents. *Journal of Consumer Psychology, 13*(1),75–91.

Häubl, G., & Trifts, V. (2000). Consumer decision making in online shopping environments: The effects of interactive decision aids. *Marketing Science, 19*(1), 4–21.

Kakalik, J., & Wright, M. (1996, Fall). Responding to privacy concerns of consumers. *Review of Business, 18*(1), 15–18.

Microsoft bCentral (n.d.). Retrieved April 3, 2003, from http://www.bcentral.com/articles

Oliver, R. W. (2001, May/June). 2B or not 2B. *The Journal of Business Strategy, 22*(3), 7–10.

Prabhaker, P. (2000). Who owns the online consumer? *Journal of Consumer Marketing, 17*(2/3),158–172.

Raisch, W. D. (2001). *The e-marketplace: Strategies for success in B2B e-commerce*. New York: McGraw–Hill.

Scott, C., & Shermach, K. (1998, September). A new era in database marketing. *Credit Card Management, 11*(6), 79–80

Spiegel, R. (2000, February 8). E-marketplaces to drive online B2B purchasing. *E-Commerce Times*.

Turban, E. (2002). *Electronic commerce 2002: A managerial perspective* (2nd ed.). Prentice Hall.

Weinberg, N. (2001, September 10). B2B grows up. *Forbes, 168,* 18–21

White Paper: Business-to-business e-marketplaces: collaboration & e-commerce (n.d.). Retrieved December 1, 2002, from http://www.Commerceone.com

Encryption

Ari Juels, *RSA Laboratories*

INTRODUCTION

Encryption is the procedure of rendering a message into a concealed form so that it is decipherable exclusively by a particular recipient or recipients. The message in its original state is known as a *plaintext* (or *cleartext*); in its encrypted form, it is known as a *ciphertext*. Historically, the aim of encryption has been to enable two parties to exchange messages confidentially, even in the presence of an eavesdropper capable of intercepting most or all of their communications. The use of encryption has been confined chiefly to diplomatic and military circles in the past, but its scope in everyday life has broadened enormously in recent years. Thanks to the rise of the Internet, it is estimated that over half a billion personal computers are equipped today with strong encryption capabilities in their Web browsing software. This includes nearly every new computer sold today.

Active users of the Internet employ encryption on a regular basis. When accepting credit card information or processing other financial transactions, most Web servers initiate encryption sessions with clients. The form of encryption used to support sessions of this kind on the Web is very strong—so strong that it is generally believed to be effectively unbreakable by even the most powerful existing computers. In most browsers, the appearance of an icon representing a closed padlock on the bottom of the screen indicates the use of encryption in a protocol known as SSL (secure sockets layer). By clicking on this padlock, a user can learn detailed information about the encryption session, much of which is explained in further depth in this chapter.

For the reader interested in a cursory introduction to encryption, without much of the detail provided in this chapter, it is possible to read the following sections as a more-or-less self-contained exposition: Some Basics, Symmetric-Key Encryption Today, the opening paragraphs of Public-Key Cryptography, The RSA Cryptosystem, and How Public-Key Encryption Is Used.

Some Basics

The science of constructing encryption algorithms and related systems is known as *cryptography*. That of analyzing and attempting to find weaknesses in encryption algorithms is called *cryptanalysis*. Together, the two complementary sciences are known as *cryptology*. Cryptologists like to explain their ideas in terms of a small troupe of fictional characters. Traditionally, Alice and Bob are the names assigned to the fictional parties wishing to exchange confidential messages with one another. The hypothetical eavesdropper on their communication is called Eve. We follow this nomenclature in our explanations in this chapter.

The operational basis of an encryption algorithm, or cipher, is a piece of information known as a key. A key serves as input to the encryption process that Alice and Bob have agreed to use. The encryption process consists of a series of instructions on how Alice, for instance, should convert a plaintext into a ciphertext. The key serves as a parameter guiding the instructions. The reverse process, whereby Bob converts a ciphertext back into a plaintext, is also guided by a key, one that may or may not be the same as Alice's. The security of traditional ciphers, that is, the privacy of the messages they encrypt, depends upon a shared key that is kept secret. For example, in one form of folklore encryption, Alice and Bob each have copies of the same edition of a particular novel. They share the identity of this novel as a secret between them. Alice encrypts a plaintext by finding a random example of each letter of her message in the novel. She writes down the page, line, and ordinal position of each of these letters in turn. The result of this process constitutes the ciphertext. Bob, of course, can reverse the process and obtain the original plaintext by referring to his copy of the shared novel. In this case, the novel itself serves as a key—one that is quite long, of course, running as it usually does to hundreds of pages. In the forms of cryptography used on computers today, the key is much shorter, typically the equivalent in length of several words or sentences.

Symmetric-Key Encryption: Introduction

When Alice and Bob make use of the same key for encryption and decryption, as in our example involving the novel, this is referred to as symmetric or symmetric-key encryption. Let us consider another folklore cipher, in which each letter of a message is replaced with another letter according to a fixed set of random, predetermined assignments. For example, the message, "MEET ME UNDER THE BRIDGE" might be encrypted as:

ZKKO ZK BWIKQ OPK UQMIFK.

This form of encryption is known as a substitution cipher. It is in fact quite easy for a skilled cryptanalyst to break. (Edgar Allen Poe, for instance, challenged readers of a newspaper in 1839 to submit English-language ciphertexts produced by letter substitution. He published the plaintexts of many such challenge ciphertexts in subsequent numbers of the newspaper.) Knowing, for example, that 'e' is the most common letter in the English language, one would be tempted—quite correctly—to identify the letter 'K', which appears most frequently in the above ciphertext, with the letter 'E' in the plaintext. More sophisticated cryptanalytic techniques for attacking substitution ciphers focus on the frequency statistics not just for single letters, but also for letter pairs and triples.

Knowing that their cipher is subject to cryptanalytic attacks of this kind, Alice and Bob might be tempted to use a different, perhaps more complex cipher, and to hide from Eve not just their key, but also the workings of the cipher. This is equivalent in effect to making the choice of the cipher a part of the key itself. An important principle enunciated by the 19th century cryptologist Kerckhoffs discourages this approach. Adhered to by contemporary cryptologists, this principle may be stated as follows: The security of a cipher should reside in the key alone, not in the secrecy of the process of encryption. The motivations behind Kerckhoffs principle are several. First of all, widespread use of a good cipher requires that its workings be divulged in some form. Even if a cipher is only disseminated through software, for example, the underlying instructions can be reverse-engineered. Thus, it is fair to assume that an attacker can learn the mechanics of the cipher. Moreover, despite the oft-demonstrated inventiveness of the cryptographic community, there is time enough to devise and refine only a limited number of basic techniques for strong new ciphers. A poorly designed cipher, even when its workings are hidden from view, is vulnerable to attack by means of an arsenal of analytic techniques refined by the cryptanalytic community over many years. These techniques are roughly analogous in spirit to the idea that leads to the discovery of the letter 'E' in the example ciphertext above, but rely on more sophisticated forms of statistical analysis.

The basic unit of information in the computer and the fundamental unit for the encoding of digital messages is not the letter, but the bit. For this reason, contemporary symmetric-key ciphers operate through the manipulation of numerical units, rather than lexicographic units. One of the earliest ciphers designed from this perspective is known as the one-time pad. Invented during World War I, the one-time pad is also of interest as the only cipher whose security is provable in the strictest mathematical sense. Formal understanding of its properties emerged in 1948–1949, with the publication of seminal work by Claude Shannon. (The security analysis of other ciphers, as we shall explain, has a strong, but less rigorous mathematical basis.)

A one-time pad is a key shared by Alice and Bob consisting of a perfectly random string of bits as long as the message that Alice wishes to transmit to Bob. To encrypt her message, Alice aligns the pad with the message so that there is a one-to-one correspondence between the bits in both. Where the pad contains a '1', Alice flips the corresponding bit in her plaintext. She leaves the other bits of the plaintext unchanged. This simple process yields the ciphertext. As may be proven mathematically—and perhaps grasped intuitively—this ciphertext is indistinguishable to Eve from a completely random string. Indeed, from her perspective, it is a completely random string. This is to say that Eve can learn no information at all from the ciphertext, no matter how powerful her cryptanalytic capabilities. Apart from its requirement of perfect randomness, though, the one-time pad carries another strong caveat. The term "one-time" refers to the fact that if Alice uses the same key to encrypt more than one message, she loses the security properties of the cipher. This makes the one-time pad impractical for most purposes, as it requires that Alice and Bob generate, exchange, and store many random bits in advance of their communications. Nonetheless, the one-time pad has seen practical use. For example, the "hot-line" established between the United States and Soviet Union in the wake of the Cuban missile crisis employed a one-time pad system to ensure confidentiality, with tapes containing random keying material exchanged via the embassies of the two countries.

Symmetric-Key Encryption Today

The symmetric-key ciphers in common use today are designed to employ relatively short keys, typically 128 bits in length. Moreover, these ciphers retain their security properties even when individual keys are used to encrypt many messages over long periods of time. Indeed, a well-designed symmetric-key cipher should permit only one effective avenue of attack, described by cryptographers as exhaustive search or brute force. By this, it is meant that an attacker familiar with the cipher makes random guesses at the key until successful. If Eve wishes to mount a brute-force attack against a ciphertext sent by Alice to Bob, she will repeatedly guess their shared key and try to decrypt the message, until she obtains the correct plaintext.

Given use of a 128-bit-long key, Eve will on average have to make well more than a trillion trillion trillion guesses before she is successful! This is more, for example, than the total number of atoms composing all of the human beings in the world. The most powerful computers available today could not be expected to mount a successful brute-force attack against a well-designed cipher employing a key of this length, even over the course of many years. It should be observed that the difficulty of breaking a key doubles for every additional bit in length. Thus, for instance, a 128-bit key is not twice as hard to break as a 64-bit key, but over a million trillion times harder.

The first widely embraced cipher employing the strong design principles in use today was the Data Encryption Standard or DES (pronounced "dehz"). Developed at

IBM, the DES cipher was published as a federal standard in 1976 by what is now the National Institute of Standards (NIST) of the United States government. DES and security-enhanced variants are still widely deployed, particularly in the banking industry. DES employs a 56-bit key, operating on a basic unit of encryption consisting of a 64-bit block. This is to say that in order to encrypt a long message using DES, Alice first subdivides the message into 64-bit (i.e., eight-byte) blocks, each of which she enciphers individually. Ciphers that operate in this fashion are referred to as block ciphers.

Brute-force attack on a 56-bit key requires substantial computational effort, very likely beyond the reach of most organizations at the time of invention of DES. Today, however, such capability is attainable with networks of ordinary workstations. This was first demonstrated in 1997 when a successful attack was mounted against a DES-encrypted ciphertext by a network of thousands of computers over the course of 39 days, and subsequently duplicated by a single, special-purpose computer in less than a day. Earlier concerns about the strength of DES had already prompted many organizations to employ a strengthened version involving application of DES operations not once, but three times to each input block, and using up to three distinct DES keys. Known as triple-DES, this enhanced version offers considerably stronger security than DES, with what may be viewed as an effective key strength of up to 112 bits.

With DES in its basic form approaching the end of its serviceable lifetime, the cryptographic community began in 1999 to lend its efforts to the development of a new standard cipher to serve as a successor. The Advanced Encryption Standard or AES (pronounced letter by letter) emerged as the result of an open competition conducted by NIST. After a period of rigorous scrutiny by the research community and government agencies, a cipher known as Rijndael (of which one recognized pronunciation is "Rhine dahl") was selected as the AES. Designed by two Belgian researchers, Joan Daemen and Vincent Rijmen, the AES promises to see widespread deployment in the United States and internationally in coming years. Rijndael is a block cipher designed to accommodate key lengths of 128, 196, or 256 bits and operate on data blocks of 128, 196, or 256 in any of the nine possible combinations. Rijndael, like many contemporary symmetric-key ciphers, is capable of very fast encryption—substantially faster than triple-DES. On a Pentium III running at 1 GHz, implementations of AES with 128-bit keys are capable of achieving an encryption speed of 50 megabytes per second. The code sizes for such implementations can be less than one kilobyte. The size of a basic ciphertext yielded by AES, as with any block cipher, is roughly the size of the plaintext. A little extra space—less than the block size—may also be needed to accommodate the fact the plaintext generally cannot be divided into data blocks of exactly the right size. See the NIST AES Web site for further information and links (National Institute of Standards and Technology, 2002).

Another symmetric-key cipher deserving discussion is RC4. Known as a stream cipher, RC4 does not operate by encrypting individual blocks of data. Rather, the only input to RC4 is a key, typically 128 bits in length. The output of this cipher is a string of random-looking bits. This string may be made as long as desired by the user. In order to encrypt a message for Bob, Alice inputs a shared key to RC4 and generates a string as long as the message. Although not a one-time pad, this string is used by Alice to encrypt the message in exactly the same manner as a one-time pad, i.e., using the same system of bit alignment and flipping. Additionally, the output string of RC4 for a particular key has the "one-time" restriction, which is to say that the string can be safely used for encryption only once. The fact that RC4 can generate a string of arbitrary length, however, means that different portions of the string can be used for different messages. Also, encipherment under RC4 is naturally capable of yielding a ciphertext identical in length to the plaintext.

This said, the output string of RC4 on a given key is not in fact a one-time pad, because it is not fully random. This may be seen in the fact that Alice and Bob know the process that generated the string, because they know their shared input key to RC4. In particular, they can write down a short set of instructions describing to someone else how to generate exactly the same output string from RC4. If k denotes their shared key, these instructions would simply say, "Give the key k as input to RC4." They could not do this in the case of a truly random string. Suppose, for instance, that Alice and Bob generated a shared random string, i.e., a one-time pad, by flipping a coin many times. If they gave the instructions "Flip a coin" to Carol, it is almost certain that Carol, in following the same instructions, would generate a very different-looking string. Obviously, though, RC4 is much more convenient for Alice and Bob to use than a one-time pad, because it only requires of them that they share a key of, say, 128 bits in length. For all intents and purposes, this is true no matter how long the messages they wish to encrypt.

The security of the RC4 cipher comes from the fact that from the perspective of Eve, who does not know the key, the output of RC4 is indistinguishable from a one-time pad. This is to say that if Alice were to give an RC4 output string to Eve (rather than using it for encipherment), and were also to give Eve a truly random string of the same length, Eve would not be able to tell the difference. Thus, as far as Eve is concerned, when Alice and Bob encrypt their messages using RC4, they might as well be using a one-time pad. This, at any rate, is a rough expression of the conclusion that cryptanalysts have arrived at after many years of statistical study of the RC4 cipher. They express this belief by describing the output of RC4 as being pseudo-random. Strong block ciphers like AES are also believed to possess this property of pseudo-randomness, but in a different form.

RC4 is of particular interest because it is one of the most commonly used ciphers in the world and included in the software in nearly every new PC sold today. It is generally a component of the SSL encryption system used for secure credit-card transactions on the Internet. In other words, the closed padlock on a browser screen mentioned earlier in this chapter indicates that the RC4 cipher is being used. RC4 was designed by Ron Rivest, one of the co-inventors of the RSA cryptosystem, which we discuss later in this chapter. The letters "RC" stand for "Ron's

Code," and the number '4' denotes Rivest's fourth cipher design. The design principal for RC4, like that of most symmetric ciphers, is a delicate sequencing of a few very basic mathematical instructions. In RC4, the operations are the swapping of integer elements in a small array representing a permutation and addition of small integers (in fact, modular addition, an operation explained later in this chapter). Others of Rivest's suite of cipher designs are also in common use, namely RC2, which forms the basis of many e-mail encryption programs, and also RC5. Apart from DES, triple-DES, AES, and the RC series, there are a number of other popular ciphers used in various systems today. These include IDEA, CAST-128, and Blowfish, to name just a few.

More on the Security of Symmetric-Key Encryption

As already explained, cryptologists know of a complete mathematical proof of security only for the one-time pad. For DES, triple-DES, AES, RC4, and kindred ciphers, cryptographers have no such security proofs. Strong proof for these ciphers is not possible; belief in their security, however, rests on fairly well-explored mathematical foundations.

Shortly after the publication of DES, two cryptologists named Eli Biham and Adi Shamir developed a technique called differential cryptanalysis. Roughly speaking, differential cryptanalysis involves statistical analysis of a cipher based on the way particular bits change (or do not change) in output ciphertexts when bits in certain positions are flipped in input plaintexts. The technique of differential cryptanalysis helped to confirm the strength of DES and to inform the design of later ciphers.(Some time after the academic development of differential cryptanalysis, it was publicly revealed that the technique had already been familiar to the government intelligence community, and had indeed helped guide the design of DES.) Subsequently developed cryptanalytic techniques have further enhanced the collection of tools available to the cryptanalyst, influencing new cipher designs in the process.

We have thus far been describing Eve as a passive eavesdropper, attempting to decipher the messages sent between Alice to Bob by harvesting and analyzing the ciphertexts they exchange. In fact, cryptologists also consider a range of active attacks that Eve might mount. By active attacks, we mean ways in which Eve might try to tamper with or otherwise influence what messages Alice and Bob exchange, in the hope of learning additional information. For example, if Alice encrypts all of her e-mail, then Eve might send a note containing a petition to Alice, and ask Alice to forward it to her friends. If Alice sends the petition to Bob while Eve is eavesdropping, then Eve learns a ciphertext for which she herself has selected the corresponding plaintext. In other words, Eve is able to perform active experimentation on the cipher. If Alice and Bob use a cipher that is poorly designed, then Eve may be able to gain information about their key in this way, or about messages they have exchanged with one another. This type of attack is known as a chosen-plaintext attack. It is an example of one of the types of active attack against which cryptographers must ensure that their cipher designs are resistant.

Even if a cipher is well designed, however, it must still be used with great care. Consider another example. Suppose that Alice and Bob always use a strong block cipher such as AES with a random 128-bit key to encrypt their communications. They perform encryption simply by dividing their messages into blocks and encrypting each block individually under their key. But suppose further that Eve knows that Alice and Bob exchange stock tips, and that these regularly take the form of simple buy and sell orders. Eve might then pass some stock tips to Alice, such as "Buy ABC, Inc." and "Sell DEF Corp.," and suggest that these be forwarded to Bob. If Eve eavesdrops while these stock tips are forwarded, then she learns what the corresponding ciphertexts look like. Thus, if Alice later sends the message "Buy ABC, Inc." to Bob, Eve will be able to recognize the ciphertext and identify the corresponding buy or sell order. In other words, even though Alice and Bob are using a very strong cipher, Eve will be able to identify (and effectively, to decrypt) any of a small set of target messages!

The way that Alice and Bob use a block cipher is known as a mode of operation. The naïve example we just described is known as electronic code book (ECB) mode, a form whose use is avoided today in part because of the problem we have just described. To prevent Eve from learning what particular ciphertexts look like, Alice may adopt a mode of operation that involves the introduction of random bits into the message. One popular mode of operation today is known as cipher-block-chaining (CBC) mode. In this system, Alice divides her plaintext into blocks and inserts some freshly generated random bits at the beginning to serve as the first block. Alice then employs a principle of "chaining" in which the encryption of one block is affected by the encryption of the previous one, thereby causing the randomness in the first block to propagate through the ciphertext. Although the architectural motivations behind this mode are somewhat complex, its basic impact on message privacy is easy to understand. The fact that Alice introduces randomness into her ciphertexts means that the same plaintext is not encrypted twice the same way, and thus that Eve cannot trick Alice in order to recognize the stock tips or other plaintexts that Alice encrypts.

Encryption and Passwords

Symmetric-key ciphers are used not only to protect communications, but also commonly to protect files against unauthorized access. Many users rely on encryption software to protect files on their hard drives against exposure to hackers or to protect sensitive data in case of laptop theft. For these purposes, it is common for the user to employ a password. The encryption software converts this password into a key for use with a standard symmetric-key cipher such as AES. Because users typically employ passwords consisting of or closely related to words in their native languages, there is generally less randomness in the key generated by a password than in a randomly generated symmetric key. One well-known means for a hacker to attack password-encrypted files in a particular encryption system, therefore, is to compile a large lexicon of common passwords. The hacker converts each entry in the lexicon into a symmetric key in the same manner as the encryption system and then uses each such key in a brute-force attack against individual users employing that system. This is known as a dictionary attack. One way to reduce vulnerability to dictionary attacks is to use salt. This is a random string of bits generated for each password individually and combined with the password in the generation of a symmetric key. The use of salt renders dictionary attacks more difficult, as it effectively forces an attacker to recompile the base lexicon for each target password. It should be understood nevertheless that salt is a limited countermeasure, and does not compensate for poor selection of passwords.

Passwords are, of course, also the most common way for users to authenticate when logging into accounts over the Internet. In this context, however, the password is typically not used as an encryption key. Instead, the server to which the user is attempting to connect checks processed password information against a database entry for the requested account.

A Brief Historical Note

Although the history of cryptology is an exciting subject of study in its own right, it is also intimately associated with the birth of the digital computer. During WWII, the efforts of British signals intelligence to break the German Enigma cipher led to the development of mechanical devices known as “bombes,” so called because of the ticking sounds they made in testing possible cipher keys. The “bombes” and the later generation of Colossus machines arising from these efforts were important precursors of the modern computer. Moreover, the man overseeing the immensely successful Enigma break was Alan Turing, a progenitor of the field of computer science.

PUBLIC-KEY CRYPTOGRAPHY

Many years of research have led to the widespread deployment of strong symmetric-key ciphers capable of high encryption speeds. One might be tempted to believe that the basic problem of private communications has been solved and that the science of cryptology has run its course. Even if a symmetric-key cipher is unbreakable, however, there remains a fundamental problem. We have assumed in our discussion that Alice and Bob share knowledge of a secret key. The question is: How do they obtain this key to begin with?

If Alice and Bob can meet face to face, in the absence of the eavesdropper Eve, then they may generate their shared key by repeatedly flipping a coin, or Alice may simply hand Bob a key written on a piece of paper or stored on a floppy disk. What if Alice and Bob wish to communicate privately over the Internet, however, without ever meeting? Alternatively, what if a commercial site on the Web wishes to enable any customer, new or old, to submit an order and credit card number securely from anywhere in the world? The administrator of the Web site cannot possibly hope to communicate keys in private to all customers before they log in. To simplify the formidable difficulties that secure key distribution can pose, cryptographers have devised a form of mathematical magic known as public-key encryption or cryptography (also known as asymmetric-key cryptography, in contrast to symmetric-key cryptography). Using public-key cryptography, Alice and Bob can send each other encrypted messages securely, even if they have never met, and even if Eve has eavesdropped on all of their communications!

In a public-key cryptosystem, Alice possesses not one, but two keys. The first is known as her public key. Alice makes this key known to everyone; she may publish it on her Web page, in Internet directories, or in any other public place. Her second, mathematically related key is known as a private key. Alice keeps this key secret. She does not divulge it to anyone else, even people she wishes to communicate with privately. Together, the public key and private key are referred to as a key pair. Bob sends a private message to Alice by performing a computation using her public key, and perhaps a private key of his own as well.

Public-key cryptography is a powerful tool—indeed, one whose feasibility may at first seem counterintuitive. Even if Eve knows Alice and Bob’s public keys, it is possible for Alice and Bob to communicate privately using public-key cryptography. Moreover, with public-key cryptography, not only Bob, but Carol or any other party can achieve private communication with Alice over a public communication medium like the Internet.

Diffie–Hellman Key Exchange

Public-key encryption was the brainchild of Ralph Merkle, who in 1974 conceived a plausible, but somewhat impractical initial scheme. The idea saw its first practical form in 1976 in a seminal paper by Whitfield Diffie and Martin Hellman entitled “New Directions in Cryptography.” Diffie and Hellman proposed a system in which Alice can combine her private key with the public key of Bob, and vice versa, such that each of them obtains the same secret key. Eve cannot figure out this secret key, even with knowledge of the public keys of both Alice and Bob. This system has come to be known as Diffie–Hellman key exchange, abbreviated D–H.

D–H exploits a form of mathematics known as modular arithmetic. Modular arithmetic is a way of restricting the outcome of basic mathematical operations to a set of integers with an upper bound. It is familiar to many schoolchildren as “clock arithmetic.” Consider a clock on military time, by which hours are measured only in the range from zero to twenty-three, with zero corresponding to midnight and twenty-three to 11 o’clock at night. In this system, an advance of 25 hours on 3 o’clock brings us not to 28 o’clock, for example, but full circle to 4 o’clock (because $25 + 3 = 28$ and $28 - 24 = 4$). Similarly, an advance of 55 hours on 1 o’clock brings us to 8 o’clock (because $55 + 1 = 56$ and $56 - (2 \times 24) = 8$). In this case, the number 24, an upper bound on operations involving the measurement of hours, is referred to as a modulus. When a calculation involving hours on a clock yields a large number, we subtract the number 24 until we obtain an integer between 0 and 23, a process known as modular reduction. This idea can be extended to moduli of different sizes. For example, in the modulus 10, the sum of 5 and 7 would not be 12, which is larger than 10, but 2, because modular reduction yields $12 - 10 = 2$. We say that an arithmetic operation is modular when modular reduction is applied to its result. For example, modular multiplication is simply ordinary multiplication followed by modular reduction. We write mod p to denote reduction under modulus p. Thus, for example, it is easily seen that $3 \times 4 \bmod 10 = 2$.

Diffie and Hellman proposed a public-key cryptosystem based on modular multiplication, or more precisely, on modular exponentiation, i.e., the repeated application of modular multiplication. Their scheme depends for its security on the use of a modulus that is a very large number. In the systems used today, the modulus is typically an integer that is 1024 bits in length, i.e., a little more than 300 decimal digits. It is also generally prime, which is to say that apart from the number one, it is not divisible by any smaller integers. There are some additional, technical restrictions on the form of the modulus that we shall not explore here.

The security of D–H is based on the following idea. Suppose that p is a large modulus and g is an integer less than p (again, with some additional technical restrictions). Suppose that Alice selects a random integer a, also less than p. She then computes the integer $y = g^a \bmod p$ and gives the integers p, g, and y to Eve. It is believed by cryptologists that with this information alone, it is infeasible for Eve to figure out the value a. The task of figuring out a is known as the discrete logarithm problem, one that has been the subject of many years of study by mathematicians and cryptographers. Although the security of D–H is not directly based on the discrete logarithm problem, it is very closely related.

In D–H, the values p and g are standard, public values. They may be conveyed in some widely distributed piece of software, such as a browser. Alice selects a random integer a less than p as her private key and computes $y_{\text{Alice}} = g^a \bmod p$ as her public key, i.e., the key that she publishes. Bob similarly selects a random integer b, also less than p, as his own private key, and computes $y_{\text{Bob}} = g^b \bmod p$ as his public key. Using her secret key a, Alice can take the public key y_{Bob} and compute a value $k = (y_{\text{Bob}})^a \bmod p$. Bob, similarly, can compute exactly the same value k using y_{Alice} in combination with his own private key b. In particular, it is also the case that $k = (y_{\text{Alice}})^b \bmod p$ (thanks to the commutative properties of modular exponentiation). Eve, however, cannot figure out the secret k. This, at least, is the belief of cryptologists, based on the idea that Eve knows neither of the private keys a or b and on the difficulty of the discrete logarithm problem. Thus, if Alice and Bob employ the secret k as the basis for private communication using a symmetric-key cipher like AES, Eve will be unable to eavesdrop successfully on their communications. This is a capsule summary of the idea behind the Diffie–Hellman cryptosystem. There are other details involved in making D–H a secure, workable system, which we gloss over in this description. We also note in passing that an attractive feature of D–H is the fact that variants may be implemented over algebraic structures known as elliptic curves. This results in more compact key lengths and faster running times.

D–H is used, among other places, in some versions of PGP (Pretty Good Privacy), a popular piece of encryption software used to secure Internet communications such as e-mail, and available as freeware in some versions.

The RSA Cryptosystem

A year after the publication of Diffie and Hellman's key exchange system, three faculty members at M.I.T. proposed a new public-key cryptosystem. This cryptosystem is called RSA after its three inventors, Ronald Rivest, Adi Shamir, and Leonard Adleman. RSA is now the most widely deployed cryptosystem in the world (see RSA Laboratories, 2002).

Use of the RSA cryptosystem for encryption is very similar to that of D–H. One superficial difference is that in the RSA cryptosystem, Bob can send an encrypted message to Alice without having a public key of his own. In particular, Bob can encrypt a message directly under Alice's public key in such a way that Alice can decrypt the ciphertext using her private key. (Similar functionality can be achieved with D–H by having Bob generate a temporary key pair on the fly and using this as the basis for encryption of a message for Alice.) There are two rather more important differences between the two cryptosystems. The first lies in the speed of their respective operations. In RSA, encryption is a very fast operation; it generally requires less computational effort than a D–H key exchange. Decryption, however, is several times slower for RSA than key exchange in D–H. Another important difference between the two is a feature present in RSA, but absent in D–H (although realized in later variants). RSA can also be used to perform digital signing, an operation not covered in this chapter, but of central importance in cryptography.

The RSA system employs modular arithmetic, the same type of mathematical basis as for Diffie–Hellman key exchange. In D–H, the modulus is a published value that may be used by any party to construct his or her key pair. In RSA, however, every party uses a different modulus, published as part of his or her public key. Thus, if Bob wishes to encrypt a message for Alice, he uses a modulus N_{Alice} unique to the public key of Alice; if he wants to encrypt a message for Carol, he uses a different modulus, N_{Carol}. The reason for the use of different moduli is the fact that the value of the modulus in the RSA cryptosystem relates directly to that of the private key.

A modulus N in the RSA cryptosystem has a special form. It is the product of two large prime integers, generally denoted by p and q. In other words, $N = p \times q$. The pair of primes p and q are treated as private values in the RSA cryptosystem; they are used to compute the private key for the modulus N. Thus the security of RSA is related very closely to the difficulty of determining the secret primes p and q given knowledge of the modulus N. This is known as the problem of factoring and is believed to be extremely difficult when p and q are large. In typical systems today, p and q are chosen to be primes of about 512 bits in length, so that N is an integer of about 1024 bits, the same length as generally selected for a Diffie–Hellman modulus. Factoring is the only effective method known for attacking RSA when the cryptosystem is used properly.

Among its many other uses, the RSA cryptosystem (along with RC4) is used as the basis for SSL, the encryption system in most Web browsers today, and also in some versions of PGP (Pretty Good Privacy) software.

More Technical Detail on RSA

We now describe in more mathematical detail how the RSA cryptosystem works. Unable as we are to delve further into the underlying mathematics, we provide only prescriptive formulae, without explaining the rationale behind them. For further information, the reader is directed to, e.g., (Menezes, van Oorschot, & Vanstone, S. A., 1996).

To compute her public key, Alice selects two random primes, p_{Alice} and q_{Alice}, both of roughly equal length in bits, e.g., 512 bits long. The product $N_{\text{Alice}} = p_{\text{Alice}} \times q_{\text{Alice}}$ is the modulus used by Alice in her public key. She also selects a small odd integer e_{Alice}; generally, this value is more or less standardized in a given system, the integer 65535 being a common choice. An additional restriction on e_{Alice} is that it must not have a factor in common with $p_{\text{Alice}} - 1$ or $q_{\text{Alice}} - 1$, a criterion considered in the selection of primes. Together, the pair of values $(e_{\text{Alice}}, N_{\text{Alice}})$ constitutes Alice's public key. Her private key is computed as follows. Let $\Phi(N_{\text{Alice}}) = (p_{\text{Alice}} - 1) \times (q_{\text{Alice}} - 1)$. Alice computes her private key, generally denoted by d_{Alice}, in such a way that $e_{\text{Alice}} d_{\text{Alice}} = 1 \bmod \Phi(N_{\text{Alice}})$. The bit-length of the private key d_{Alice} here is typically quite close to that of N_{Alice}.

A message in the RSA cryptosystem consists of a positive integer m less than N_{Alice}. To encrypt m, Bob computes the ciphertext $c = m^{e_{Alice}} \bmod N_{\text{Alice}}$. As a technical restriction required to achieve good security, the plaintext m is formatted so as to be about equal in length to the modulus, resulting in a ciphertext that is similarly so. To decrypt the ciphertext, Alice applies her private key and computes $m = c^{d_{Alice}} \bmod N_{\text{Alice}}$.

An Example

Let us consider a small example to provide some flavor of how the RSA cryptosystem works. For illustrative purposes, we consider integers much smaller than those needed for true, secure use of RSA. Suppose that p_{Alice} and q_{Alice} are 5 and 11 respectively. Then $N_{\text{Alice}} = 5 \times 11 = 55$. Observe that $e_{\text{Alice}} = 3$ does not divide $p_{\text{Alice}} - 1$ or $q_{\text{Alice}} - 1$. Thus, Alice can use $(e_{\text{Alice}}, N_{\text{Alice}}) = (3, 55)$ as her public key. A valid private key for Alice is $d_{\text{Alice}} = 27$, because $\Phi(N_{\text{Alice}}) = (p_{\text{Alice}} - 1) \times (q_{Alice} - 1) = (5 - 1) \times (10 - 1) = 40$, and $e_{\text{Alice}} \times 27 = 81 = 1 \bmod 40$.

To encrypt the message $m = 7$, Bob computes the ciphertext $c = 7^3 \bmod 55 = 343 \bmod 55 = 13$. Alice decrypts the ciphertext $c = 13$ by computing $13^{27} \bmod 55$. Using a pocket calculator, the reader can easily verify that the result of this decryption operation is indeed the original plaintext $m = 7$.

The security of RSA requires some special, technical restrictions on the value of m. In fact, as the RSA cryptosystem is generally employed, the message m itself encodes the key for a symmetric-key cryptosystem such as AES. Public-key cryptosystems are generally used in conjunction with symmetric-key cryptosystems, as we discuss below.

With a 1024-bit RSA modulus, a Pentium III running at 1 GHz can perform roughly 40 RSA decryption operations per second, and about 600 encryption operations per second under the public exponent $e = 65535$. (These speeds, of course, depend critically on the particular software implementation.)

As explained above, the security of the RSA cryptosystem is closely related to the difficulty of factoring the product of two large primes. Security guidelines for RSA thus depend upon advances in the factoring problem made by researchers, as well as the cost and availability of computing power to a potential attacker. According to guidelines issued in 2001 by NIST, 1024-bit RSA moduli may be used to secure data whose privacy needs to be assured until the year 2015. For more sensitive data, a 2048-bit RSA modulus is recommended instead. Although there is substantial debate and uncertainty among cryptographers as to the exact ongoing security level afforded by RSA, there is general agreement as to the rough accuracy of these predictive guidelines. To avoid a common misconception, it should be emphasized that the mathematical basis for the RSA cryptosystem means that the recommended modulus lengths, and thus the public and private keys, are substantially longer than the 128- to 256-bit key lengths prescribed for symmetric-key ciphers.

How Public-Key Encryption Is Used

Encryption of a large quantity of data is generally much slower using a public-key cryptosystem than using a symmetric-key cipher. On the other hand, public-key encryption offers an elegant approach to the problem of distributing keys that is unavailable in symmetric-key ciphers. It is common in practice, therefore, to combine the two types of encryption systems in order to obtain the best properties of both. This is achieved by use of a simple principle known as enveloping. Enveloping involves use of a public-key cryptosystem such as RSA as a vehicle for transporting a secret key, which is itself then used for encryption with a symmetric-key cipher. To send a megabyte-long file to Alice, for example, Bob might select a random 128-bit RC4 key k and encrypt the file under k. He would then send this ciphertext of the file to Alice, along with an RSA encryption of k under Alice's public key. In effect, enveloping is a way of making a public-key cryptosystem faster.

As with symmetric-key ciphers, cryptologists aim to design public-key cryptosystems to withstand a wide range of possible attacks, involving some very strong ones in which, e.g., Eve can persuade Alice to decrypt a range of ciphertexts that Eve herself selects. Like symmetric-key block ciphers, when used in a naïve manner, RSA has a limitation in that encryption of the same plaintext always yields the same ciphertext. Prior to encryption under RSA, therefore, it is common practice to subject a message to a special type of formatting that involves the addition of random bits, plus some additional processing. Loosely speaking, this addition of randomness serves much the same purpose as in the case of cipher-block chaining for symmetric-key block ciphers, as described above. Other aspects of the formatting process permit the RSA cryptosystem to withstand other forms of attack, and to do so with guarantees that are subject to rigorous mathematical justification.

Most public-key encryption systems have an additional advantage over symmetric-key systems, in that they permit a flexible approach to distributed storage of the private key, known as secret sharing. This is a way of mathematically splitting a private key into a number of elements known as shares. To achieve a decryption operation, each of these shares may be applied individually to a ciphertext, without the need to assemble the shares themselves in a single place. Thus, a cryptosystem can be set up so that compromise of any one share does not expose the full decryption key itself. For example, Alice can divide her private key into, say, three shares, keeping one for herself and giving one each to her friends Bob and Carol. To decrypt a ciphertext, Alice can ask Bob and Carol to apply their shares and to send her the result. Alice can then complete the decryption with her own share, without Bob or Carol seeing the resulting plaintext. If an attacker breaks into Alice's computer and obtains her share, the attacker will still be unable to decrypt ciphertexts directed to Alice (without also obtaining the assistance of Bob and Carol). The system can even be set up so that decryption is possible given particular sets of shares. Thus, for example, Alice might be able to achieve decryption of her ciphertexts even if Bob or Carol is on vacation.

Despite its great flexibility, public-key cryptography does not directly address all of the challenges of key distribution. For example, we have said that Alice can safely disseminate her public key as widely as she likes, publishing it in Internet directories and so forth. But when Bob obtains a copy of Alice's public key, how does he really know it belongs to Alice, and is not, for example, a spurious key published by Eve to entrap him? To solve problems of this kind, we appeal to a public-key infrastructure (PKI), as explained in detail in the chapter on that topic. Another issue worth mentioning—and a pitfall for many system designers—is the problem of finding an appropriate source of randomness for generating keys. Good generation of random bits is the cornerstone of cryptographic security and requires careful attention.

A final problem that encryption alone does not solve is that of message integrity. When Bob sends a message to Alice encrypted using, e.g., RC4, he may be reasonably well assured of the privacy of his communication if he uses the cipher correctly. When Alice receives his message, however, how does she know that Eve has not tampered with the message en route, by changing a few bits or words? Indeed, how does she even know that Bob is the one who sent the message? For this type of assurance, some additional cryptographic apparatus is required in the form of a message authentication code or a digital signature. These are techniques for applying a key to a message to obtain an unforgeable "fingerprint" that shows evidence of any tampering.

CONCLUSION: FURTHER READING

Two broad and accessible introductions to the field of cryptology are *Applied Cryptography*, by Bruce Schneier (2002a), and the online compendium *Frequently Asked Questions About Today's Cryptography* (RSA Laboratories, 2002). A more detailed technical treatment of cryptographic techniques and concepts, along with an extensive bibliography, may be found in the excellent *Handbook of Applied Cryptography*, by Alfred J. Menezes, Paul C. van Oorschot, and Scott A. Vanstone (Menezes *et al.*, 1996). Detailed information on the AES algorithm (Rijndael) is available in *The Design of Rijndael*, by Joan Daeman and Vincent Rijmen (2002).

Readers interested in the practical application of cryptography in real-world systems may wish to consult *Network Security: Private Communication in a Public World*, by Charles Kaufman, Radia Perlman, and Mike Speciner (2002), and *Cryptography and Network Security: Principles and Practice*, by William Stallings (1998). Also of interest to such readers may be Bruce Schneier's *Secrets and Lies* (Schneier, 2002b), which offers some caveats regarding the limitations of cryptography.

A good introductory textbook of a more academic flavor is *Cryptography: Theory and Practice*, by Douglas Stinson (2002). For information on the foundational mathematics and theory of cryptology, an important work is *Foundations of Cryptography: Basic Tools*, the first in an evolving series of volumes by Oded Goldreich (2000). For a historical overview of cryptology, the classic text is *The Codebreakers*, by David Kahn (1996). More up-to-date, although less exhaustive, is *The Science of Secrecy from Ancient Egypt to Quantum Cryptography*, by Simon Singh (2000).

Cryptologists continue to devote effort to the development of new symmetric ciphers and public-key cryptosystems, as well as the improvement of existing ones. Although encryption is the fountainhead of contemporary cryptology, it is by no means the only focus of the field. As is the case with many branches of science, cutting-edge inventions in cryptology often lie dormant for years or decades before their widespread use. The scope of cryptologic research today extends beyond the problems of message privacy and integrity to the goal of achieving fair play in electronic environments in a much broader sense. Secure electronic voting, on-line privacy protection, and digital rights management are just a few of the many areas where researchers in cryptography have made strides in recent years. At the frontiers of cryptology are ideas involving the use of quantum mechanics and even DNA for attacking ciphers. Many cryptologists are members of the International Association for Cryptologic Research (IACR), whose home page (IACR, 2002) lists publications and conferences devoted to advanced current research in cryptography and cryptanalysis.

GLOSSARY

AES The Advanced Encryption Standard, a symmetric-key cipher known as Rijndael, serving as successor to DES.

Asymmetric In the context of encryption, a type of cryptographic system in which a participant publishes an encryption key and keeps private a separate decryption key. These keys are respectively referred to as public and private. RSA and D–H are examples of asymmetric systems. *Asymmetric* is synonymous with *public-key*.

Ciphertext The data conveying an encrypted message.

Cryptanalysis The science of analyzing weaknesses in cryptographic systems.

Cryptography The science of constructing mathematical systems for securing data.

Cryptology The combination of the complementary sciences of cryptography and cryptanalysis.

Cryptosystem A complete system of encryption and decryption, typically used to describe a public-key cryptographic system.

Decryption The process of obtaining a readable message (a *plaintext*) from an encrypted transformation of the message (a *ciphertext*).

DES The Data Encryption Standard, an existing form of symmetric cipher in wide use today, often in a strengthened variant known as triple DES.

Diffie–Hellman (D–H) A public-key cryptosystem used to exchange a secret (symmetric) key.

Encryption The process of rendering a message (a *plaintext*) into a data string (a *ciphertext*) with the aim of transmitting it privately in a potentially hostile environment.

Enveloping A method for using a symmetric-key cipher in combination with a public-key cryptosystem to exploit simultaneously the advantages of the two respective systems.

Key A short data string parameterizing the operations within a cipher or cryptosystem, and whose distribution determines relationships of privacy and integrity among communicating parties.

Key pair The combination of a public and private key.

Plaintext A message in readable form, prior to encryption or subsequent to successful decryption.

Private key In an asymmetric or public-key cryptosystem, the key that a communicating party holds privately and uses for decryption or completion of a key exchange.

Public key In an asymmetric or public-key cryptosystem, the key that a communicating party disseminates publicly.

Public-key In the context of encryption, a type of cryptographic system in which a participant publishes an encryption key and keeps private a separate decryption key. These keys are respectively referred to as public and private. RSA and D–H are examples of public-key systems. *Public-key* is synonymous with *asymmetric*.

RC4 A symmetric-key cipher of the type known as a stream cipher. Used widely in the SSL (secure sockets layer) protocol.

RSA A public-key cryptosystem in very wide use today, as in the SSL (secure sockets layer) protocol. RSA can also be used to create and verify digital signatures.

Symmetric A type of cryptographic system in which communicating parties employ shared secret keys. The term is also used to refer to the keys employed in such a system.

CROSS REFERENCES

See *Guidelines for a Comprehensive Security System; Internet Security Standards; Passwords; Public Key Infrastructure (PKI); Secure Electronic Transmissions (SET).*

REFERENCES

Daeman, J., & Rijmen, V. (2002). *The Design of Rijndael.* Berlin/New York: Springer-Verlag.

Goldreich, O. (2000). *Foundations of Cryptography: Basic Tools.* Cambridge, UK: Cambridge University Press.

Kahn, D. (1996). *The Codebreakers: The Story of Secret Writing.* New York: Simon & Schuster.

Kaufman, C., Perlman, R., and Speciner, M. (2002) *Network Security: Private Communication in a Public World* (2nd ed.). Englewood Cliffs, NJ: Prentice–Hall.

International Association for Cryptologic Research (IACR) home page (2002). Retrieved 2002 from http://www.iacr.org

Menezes, A. J., van Oorschot, P. C., & Vanstone, S. A. (1996). *Handbook of Applied Cryptography.* Boca Raton, FL: CRC Press.

National Institute of Standards and Technology (NIST) Web page on the AES Algorithm (Rijndael) (2002). Retrieved 2002 from http://csrc.nist.gov/encryption/aes/rijndael

RSA Laboratories (2002). *Frequently asked questions about today's cryptography*. Retrieved 2002 from http://rsasecurity.com/rsalabs/faq/index.html

Schneier, B. (2002a). *Applied cryptography: Protocols, Algorithms, and Source Code in C* (2nd ed). New York: Wiley.

Schneier, B. (2002b). *Secrets and lies: Digital Security in a Networked World.* New York: Wiley.

Singh, S. (2000). *The Science of Secrecy from Ancient Egypt to Quantum Cryptography.* New York: Alfred A. Knopf.

Stallings, W. (1998). *Cryptography and Network Security: Principles and practice*. Englewood Cliffs, NJ: Prentice Hall.

Stinson, D. (2002). *Cryptography: Theory and Practice* (2nd ed.). Boca Raton, FL: CRC/C&H.

Enhanced TV

Jim Krause, *Indiana University*

INTRODUCTION TO ENHANCED TV

Background

For years film and television producers have been operating within the framework of broadcast. In the broadcast model, a signal is sent out from one or more points to reach a great number of installed receivers. Viewing is a passive experience, with options limited to changing channels or switching off the set. Through technological advances, the medium has evolved, providing more numerous channels and methods of delivery, higher quality sound and images, and the addition of embedded information to the video signal. Just recently the paradigm of one-way communication has begun to crumble as broadcasters and cable companies experiment with and implement two-way or interactive communication. Interactive television allows viewers to customize their television experience. They can choose what programs or information to watch and when to watch it and have the ability to use chat and e-mail, order products, and make financial transactions.

Enhanced TV bridges the gap between the historical broadcast model and interactive communication.

eTV or iTV?

In 1977 Warner Cable launched the first interactive television system by introducing Qube to some 30,000 viewers in Columbus, Ohio. Viewers using the experimental system could select TV programs, participate in live polls, and purchase pay-per-view movies. Despite attempts to expand service into Cincinnati and Dallas, Qube ultimately closed down due to prohibitively high system expense. About the time of Qube's demise, the 1984 Cable Act was accelerating the growth of cable. By the end of the decade cable reached into more than 50 million homes. Digital cable and digital broadcast satellite growth continued strongly through the end of the 20th century. Perhaps due to the concurrent growth of the Internet, unsuccessful early implementation, and a shaky investment history, interactive television had acquired, at least to some, a bad reputation.

Some view enhanced television as a subset of interactive television (iTV). Others consider it as a more favorable term for interactive television, clean of any stigma that may have been attached to it. Regardless, iTV is an umbrella term that can apply to a broad range of services relying on two-way communication between the viewer and content provider (CRTC, 2002). Enhanced television shares some key features with interactive television, such as a software interface and the ability in some systems for the user to transmit data, so it's easy to see why the two are sometimes confused. To clarify, a fundamental precept behind interactive television is that the user can influence the content or storyline. Enhanced TV is primarily concerned with augmenting the content of programming, not shaping it.

This chapter addresses the following:

What is enhanced television?

What infrastructure supports enhanced television?

What is the impact of enhanced television?

What business models can be applied to enhanced television?

Definitions

Microsoft sums it up: "Enhanced TV is any programming that enables consumers to interact directly with the show they're watching" (Microsoft TV, 2002). As enhanced television can describe a broad scope of interactive technologies, two subcategories can be defined. The first is rooted in the historical broadcast model while the second refers to a system of two-way communication. Both systems require set-top-boxes or other decoding devices.

1. "Basic" enhanced TV/one-way interactive—This system offers the viewer metadata, which is additional

information augmenting the program's content. While there is no return path or back channel, viewers can view additional data or make selections that display information on the program itself or regarding selected items. Programming enhanced in this manner can be transmitted over analog or digital means, as the information can be delivered via the vertical blanking interval (VBI) of the television signal.

2. "Interactive" enhanced TV/two-way interactive—A key component of this system, which is popular in the UK, is a return path or back channel. Viewers' selections and potentially other data (answers to questions, credit card numbers, etc.) can be sent back to the place of signal origination, or to anywhere else connected to the network: marketing agencies, advertisers, or content providers. As in the first category, content providers can use analog or digital transmission, delivering additional information in the VBI, but there must be some method of a return path.

Subsets and variations of each of these two types of enhanced television will be explored later in this chapter.

Convergence Factor

The convergence of broadcast, cable, telephone, satellite and Internet-based applications presents a dizzying array of new opportunities to broadcasters, cable companies, and viewers. Once considered television providers, satellite and cable companies are now dishing out broadband Internet service. Telco/IP networks can provide television programming. Microsoft, the company that dominates the PC software industry, is deeply entrenched in television and cable networks and manufactures set-top-boxes and the software that resides inside. People can listen to the radio and watch TV on computers, and surf the Web on TVs.

Users have grown accustomed to and are demanding more information and more data. Since the introduction of the Internet, Web surfers have learned to click a link to find out more, to order products, to download files, or to play games and interact with others. From Web browsing, users have grown accustomed to clicking a link to trigger pop-up video clips and instantly jump to a new URL. In this manner, enhanced television can be interpreted as an outgrowth of the Internet, even though its roots are planted firmly in the blue glow of the family TV set. Formerly WebTV, Microsoft MSN TV's set-top-box offers users complete access to the Internet through their television sets. The lines are becoming blurred between different types of media. Users can receive cable broadcasts on the same lines that supply broadband Internet connections.

MTV: The Proverbial Yardstick

In 1981, MTV was in the spotlight and music videos became the latest novelty. Just watching Michael Jackson's moonwalk and videos from bands like the Police used to be entertainment. Now some 20 years later, viewers need more to keep them tuned in. MTV has reduced the amount of time dedicated to music videos and resorted to varied programming lineup including news, reality and game shows, documentaries, and even wrestling. CMT, VH1, and MTV augment their music videos with additional information. Through programs like *Pop-Up Video*, viewers learn that the waitress in a music video would later become a famous actress or that the red guitar once belonged to a Beatle. Some may find it interesting that VH1's popular series of documentaries, *Behind the Music*, provides viewers with information on popular musicians whose careers were fueled in part by being featured on MTV. The tendency to insert additional information into program content is becoming more popular with other networks as well. On the American Movie Classics (AMC) program *DVD-TV*, movies are offered with facts and trivia, presented at the bottom of the screen, related to the talent or production. The programming on CNN Headline News and CNBC networks resembles an enhanced television display. The primary image is squeezed into a corner, while weather reports, stock quotes, and news bulletins border the edges. So even without interactive set-top-boxes, stations are trying to squeeze as much varied and additional information as possible into their programming.

APPLICATIONS OF ENHANCED TV

Definitions aside, the real question is what can it do. What does enhanced TV look like? Program providers can provide enhanced content through single- or dual-screen applications. Methods of the specific relationships program providers can have with their audience have also been identified. Technically, one can categorize types of enhanced services according to the method of delivery, the corresponding interface and equipment required, and the means (if any) of returning information from the viewer.

Two- and Single-Screen Applications

Two-screen systems refer to a TV and the monitor of a computer connected to the Internet (see Figure 1). Single-screen systems rely on set-top-boxes or other decoders that can superimpose information directly on top of a television display. A major advantage of two-screen systems is that there is a huge base of potential users—households with televisions and computers with Internet access in the same room. People who surf the Web while watching TV are sometimes called "telewebbers" (Davis, 2000). In April of 2001 estimates counted approximately 20 million households with PCs in the same room as television sets (Zdnet News, 2001). One year later, *USA Today* estimates showed between 30 and 40 million people watch TV and view the Web simultaneously, compared to approximately 10 million homes with interactive-capable satellite or cable service (*USA Today*, 2002).

Two-Screen Systems

In lieu of superimposing additional information over the TV broadcasts, which can only be accessed by viewers with service plans and set-top-boxes or other decoders, broadcasters can provide time synchronous, program-related content over the Internet. (Truly dedicated enhanced content providers do both!) This provides broadcasters with a relatively straightforward manner in

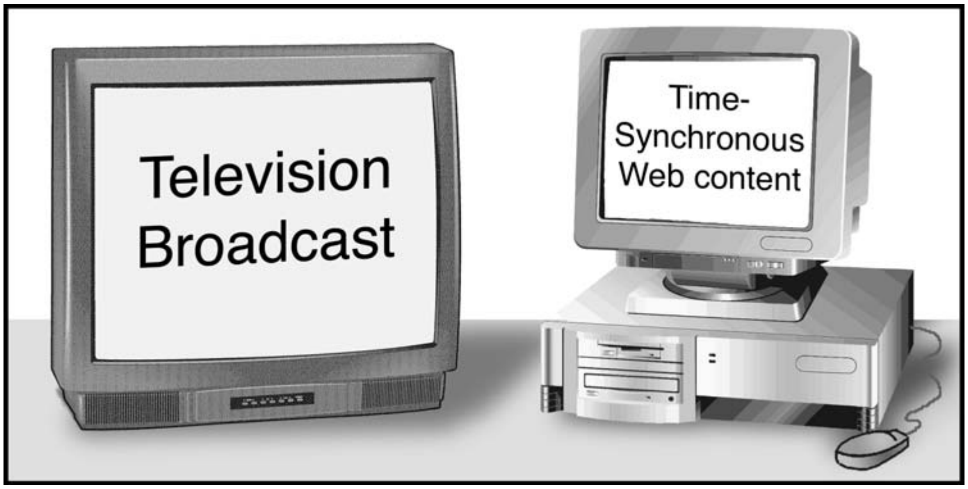


Figure 1: Two-screen application. From http://www.ttop.com. Copyright by Tabletop Productions. Reprinted with permission.

which to reach and maintain their viewers' interest in the programming outside of or even during the broadcast.

Although offering little or no embedded metadata in its programming, ABC touts its enhanced television programs on the company's Web site. Viewers watching shows like *Alias*, *Who Wants to Be a Millionaire?*, and *Sunday Night Football* can find a number of interesting ways to play along with their favorite shows. ABC's eTV Web site offers viewers a chance to participate in polls, answer trivia questions, build fantasy football teams, play interactive games, and even win prizes.

Web Strategies. In a study of enhanced television Web sites, three strategic models have been identified according to intended target audiences and strategic functions (Ha, 2002): "welcome all," "fans-friendly," and "hello." "Welcome all" Web sites are targeted equally to viewers and nonviewers. Web visitors can find general information on services and the station's programming. The "fans-friendly" model is targeted specifically to TV viewers. Web visitors who may have just wandered by might feel like they're missing out on something or a bit disoriented. "Hello" model Web sites are geared toward those with no knowledge of the station's programming.

ABC's eTV Web site could fit the "fans-friendly" model. It offers little information regarding the station's nonenhanced television programming, but offers direct links to upcoming interactive events. New users can register, and a tour button that promises to show how users can interact with their favorite programs is prominently displayed.

Single-Screen Systems

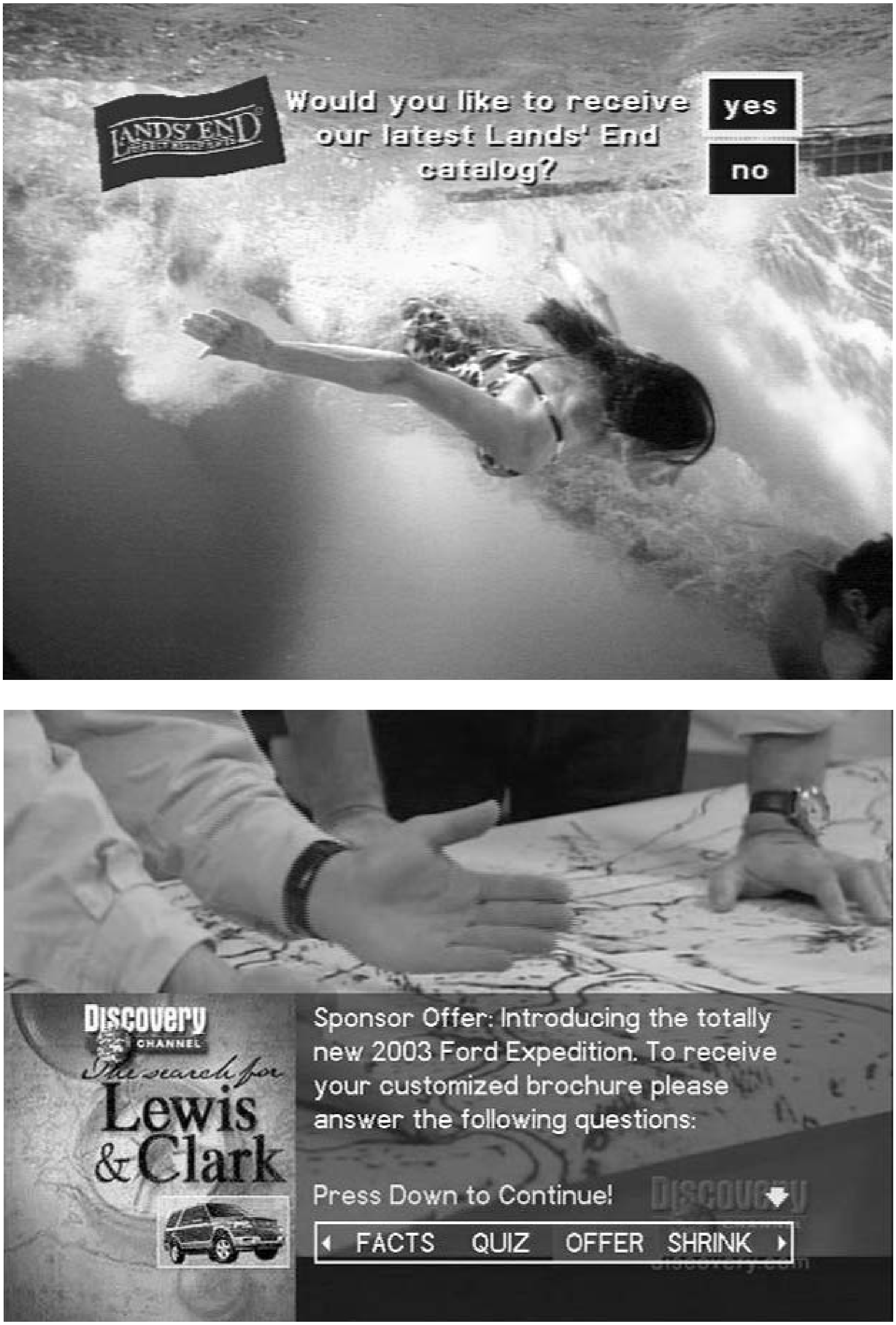
CBS was one of the first networks to bring single-screen interactive content to prime time shows. Viewers of *C.S.I.* (*Crime Scene Investigation*) with either MSN TV or Ultimate TV boxes can view case files, visit the crime lab, see short bios on the characters, and even track clues to help solve the case. *Survivor Africa* lets viewers see the voting statistics, contestant bios, and information on game challenges and history.

Single-screen systems require the use of set-top-boxes or computers with decoder cards. The additional data are superimposed over the television signal, or in some cases, the program signal is reduced in size to make room for the other content. Single-screen two-way systems also allow users to access related Web pages without entering a URL. Through what is sometimes called TV or channel hyperlinking, viewers can be easily redirected to Web sites or be provided with additional on-screen information. The latest crop of multimedia PCs from Hewlett–Packard and Sony feature tuner cards, providing access to enhanced broadcasts without a set-top-box. Because of the limited size of television and video monitors, and the fact that interfaces share the screen with the program content, the design must be simple, yet aesthetically pleasing.

Wink. Wink users can shop and view additional sports, weather, and news information through cable networks including Discovery, TLC, TechTV, The Weather Channel, ESPN, CNN, CNN Headline News, ABC, NBC, CNBC, FOX, CNBC, TBS Superstation, TNT, USA Network, and E! Entertainment. To use Wink, viewers first must register their credit card and profile into the system. Wink transmits its data using the VBI. Because of the limited amount of room available in the VBI, Wink's pop-up displays tend to be simple and straightforward. When viewers see the Wink logo on their screen, they can click their remotes to display on-screen information and interactive content. Wink TV reaches more than five millions homes and is available on DirectTV and a number of cable systems including Comcast, AT&T, Charter, Time Warner, Rogers, and Adelphia. For two samples of Wink screenshots, see Figure 2.

Types of Relationships

The types of relationships viewers have with enhanced television content can be categorized into four areas (Hurst, 2000): fan-based, game-based, information-based, and programming-based. Fan-based relationships focus on the viewers' desire to get closer to the star. Game-based enhanced content tries to recreate the fun and excitement



Would you like to receive our latest Lands' End catalog?

Sponsor Offer: Introducing the totally new 2003 Ford Expedition. To receive your customized brochure please answer the following questions:

Press Down to Continue!

Figure 2: Wink screenshots from http://www.wink.ocm. Copyright by Wink Communications. Reprinted with permission.

of the particular game. For instance, viewers can play along with *Jeopardy* or *Who Wants to be a Millionaire?* ABC's enhanced TV Web site gives sports fans the ability to assemble their own fantasy teams and call plays during the televised games. Information-based refers to the viewers' need for extended content, often in the areas of news, weather, and sports. Programming-based interactivity delivers programs customized for the viewer's personal tastes. This is achieved through personalized custom channel selections and personal video recorders, allowing the user to record programming so that they can play it back whenever they desire, without commercial interruption.

MODELS OF ENHANCED TV

In order to implement enhanced TV, a basic system with certain components must be in place: an interactive TV software application, a means of introducing additional content into the application, and a method of delivering content to eTV-enabled receivers or set-top devices. The following are operational models of a few of the more common types of enhanced TV systems, along with the basic system requirements.

System Design

Methods of Propagation

Because program metadata can be packaged within the vertical blanking interval, one-way, interactive enhanced content can be delivered to the viewer in digital or analog form through cable, terrestrial on-air broadcast, the Internet, and satellite. The return channel can use any form of Internet connectivity including cable, DSL, telephone service, and even satellite uplinking.

Terrestrial

Broadcasters in the United States are in the process of upgrading to DTV transmission. The new digital standards will allow for increased resolution, HDTV, and expanded data transmission content. At first glance, terrestrial-based digital broadcasting's main obstacle is the obvious lack of a return channel. Although this doesn't affect one-way enhanced analog or digital broadcasts, it presents an impediment to two-way interactivity. Still, terrestrial broadcasters have a few options. Alternative return channels can be routed through Telco/IP systems, as demonstrated by the satellite industry, as well as wireless return channels, providing the right equipment is in place.

After success with its digital terrestrial broadcasting standard, the European consortium DVB (see the following section Setting Future Standards: Development Communities) has been working to establish a wireless return channel. This new standard has been outlined by DVB-RCT (Digital Video Broadcast Return Channel Terrestrial), which provides a return channel for DVB-T (Faria, 2002).

Satellite

In the mid 1990s, satellite digital broadcast systems (DBS) first became available to the public. At the time of this chapter (December 2002), DirecTV and DISH Network are the largest digital satellite providers with nearly 20 million installed systems. Their digital systems provide one-way digital interactivity and two-way interactivity employed through land-based return channels. If a customer orders a pay-per-view (PPV) through DISH Network, the set-top-box (STB) uses the telephone line to bill the user's account. Hughes Network Systems developed a system allowing users to download data from a satellite and upload over a telephone modem. Their newer development, DirecPC, is a two-way system allowing downloading and uploading from a satellite. Starband is another company offering bidirectional satellite networking.

Cable

The Cable Act of 1984 helped deregulate the industry and gave a strong boost to cable growth. The 1990s saw many cable companies upgrading their systems to two-way digital networks, capable of supporting a wide range of services. According to the National Cable and Telecommunications Association (NCTA, n.d.), in 2001 digital cable service and cable modems were available to more than 65 million homes. Digital cable households can receive standard cable, digital cable, telecommunications services, and high-speed Internet access. Internet users tired of slow dial-up connections have increasingly turned to digital cable for its "always on" broadband Internet service. The scope of two-way interactivity varies with the cable distributor as well as the geographic location.

Telco/IP Systems

In addition to terrestrial, cable, and satellite, IP systems are a viable means to bring television into a household and provide users a means to receive interactive content. While telephone companies have a long history of providing Internet connectivity through dial-up connections, they have only recently begun to offer broadband service at rates comparable to cable. IP television providers across the globe have been experimenting with delivering programs and other full motion video. Broadband IP networks have been designed from the ground up to support two-way communication and potentially offer a higher degree of interactivity. This two-way communication can pose a problem in the flow of communication over such networks. When demand is high, the whole network can slow down.

Setting Future Standards: Development Communities

There is a necessity for individuals, organizations, and industries with shared interests to form alliances. These can be viewed as communities in that they have organizational structure, establish self-imposed guidelines, provide a means of sharing information and interaction through forums, electronic newsletters, and conferences, and exist to protect the interests of their membership. The growth of interactive programming is particularly significant, as it affects an unprecedented wide range of industries. The careers and livelihoods of telephone technicians, software developers, computer companies, cable companies, video producers, and screenwriters will all be affected by the growth of interactive television.

Interactive digital broadcasters from around the globe participate in industry-led consortiums to ensure open standards and a means of sharing development information. These cross-industry alliances minimize investment risk and help create agreed-upon standards.

AFI

The American Film Institute serves film, television, and video artists and is dedicated to "preserving the art of the moving image" (http://www.afi.com). In 1998 The AFI began sponsoring a series of enhanced television workshops with the purpose of exploring the best ways of developing effective interactive programming compatible with the new and emerging technologies. Their ongoing workshops provide television producers a means to showcase their work and stimulate creation of working prototypes.

ITVT

Tracy Swedlow's InteractiveTV Today is a Web-based organization featuring a number of resources for iTV developers. ITVT's Web site, http://www.itvt.com, provides a newsletter, journal articles, an online glossary, and screenshots of a number of enhanced and interactive television examples.

ATVEF

The Advanced Television Enhancement Forum (ATVEF) seeks to create a unified pathway for interactive content developers ensuring widespread compatibility throughout cable, terrestrial, satellite, and local distribution networks, regardless of the distribution path. ATVEF's enhanced content specification outlines the infrastructure supporting one- and two-way interactivity. To minimize the need for developing new technologies, the ATVEF specification supports existing Web standards such as HTML, CSS, and JavaScript. The specification also allows the sharing of data between digital and analog systems, as well as networks with no video at all (ATVEF, 2002). In order to work with varying video networks standards, the ATVEF specification requires IP bindings developed for each protocol. Content specifications provide a means to incorporate HTML and TV content. The organization offers membership and news and provides technical documents from their Web site at http://www.atvef.com.

ATSC

The Advanced Television Systems Committee was formed in 1982 as a not-for-profit, cross industry organization seeking to develop voluntary standards (ATSC, n.d.). The FCC has worked closely with the organization and adopted ATSC-developed standards for the changeover to digital transmission, which is taking place across the country. Their DTV standard has also been adopted in Canada.

SMPTE

The Society of Motion Picture and Television Engineers (SMPTE) is the leading technical society of the film and video industry. Formed in 1916, SMPTE functions to establish industry standards, enhance education through seminars, communicate the latest developments in technology, and promote networking and interaction (SMPTE, 2002). SMPTE has recently created a new standard, SMPTE DDE-1, a compliant version of ATVEF 1.1r26.

DVB

Mainly composed of members from the European television, broadcast, and data industries, Digital Video Broadcasting (DVB) is an industry-led consortium focused on developing global standards for digital television and broadcast (DVB, 2002). Through various working groups, DVB facilitates the development of open standards and widespread interoperability and has already successfully implemented a series of technical standards for digital broadcast. The DVB-MHP (multimedia home platform) standard defines a layer allowing third-party applications to interface with a variety of decoding devices including STBs, integrated TV sets, and multimedia PCs (Jones, 2002). The DVB-MHP (multimedia home platform) begun in 1997 has been adopted by ETSI, the European Telecommunications Standards Institute. The DVD-MHP version 1.0.1 specifies Java as the application environment for both enhanced and interactive broadcasts and has recently been adopted by a U.S. company, CableLabs. The set-top-box manufacturer will begin selling units following the DVB-MHP specification.

Middleware

Middleware is software and hardware that helps applications work together. In the case of enhanced TV, companies like OpenTV, Liberate, and Microsoft TV create systems that support the set-top-box applications, ensuring they interface with the data systems back at the point of origination. Middleware typically provides subscriber definitions, channel assignments, packages and pricing, video asset and metadata management, and transactions and billing interfaces (Hawley, 2002). Middleware providers develop the software and interfaces that allow users to select programming and products, connect to the Internet, and make secure transactions. Current ATVEF standards support program development through a number of third parties using readily available Web development tools.

Other companies create software that users interface with through using their remote controls and set-top-boxes. Examples include the Wink interface, an interactive program guide (IPG), or the video-on-demand (VOD) interface. With ACTV's HyperTV, content creators can develop synchronized Web sites to add interactivity to programs or advertising materials targeting viewers with compatible single-screen decoding devices. HyperTV has been used to develop material for TBS, MTV2, TNT, TV Land, and Starz/Encore. The Disney Company is another interactive developer, which developed the Internet-enhanced broadcasts of ABC's *Monday Night Football*, *Who Wants to be a Millionaire?*, and ESPN's *Sunday Night Football* (Rudnick, 2000).

Set-Top-Boxes (STB)

The set-top-box is a small computer that runs software allowing users to interact with their distribution networks. It receives and decodes signals and displays them on the television monitor. STBs for one-way systems can access embedded information and superimpose it on the TV screen. Two-way systems send information back to the broadcaster. As mentioned previously, computers with the right configurations can perform the functions of a set-top-box. Digital cable STBs contain a cable modem, and can send information back to the head-end through the cable distribution network.

Satellite networks use specialized set-top-boxes that incorporate an integrated receiver decoder (IRD). IRDs receive and descramble satellite signals. Most installed satellite set-top-boxes for both analog and digital systems require a connection to a telephone line, which serves as a return channel. Users' responses to polls, interactivity with game shows, or catalog purchases are sent back through the return channel to the source of origination. The STB keeps track of PPV movies or VOD, and then transmits the information back to the satellite distributor in order to process billing information. Some nonsatellite systems, such as MSN TV (formerly WebTV), employ the same technique, using a phone line as the means to send return data. Some of the next generation of STBs will likely be able to use alternate means of sending return data through a LAN or other Internet connection.

Cable distribution companies tend to provide STBs to subscribers at no initial cost, subsidizing the expense through monthly service fees. The choice of STBs is usually limited to assure compatibility and to keep costs down. Digital satellite providers tend to sell STB/IRDs separately, offering a range of choices.

Personal Video Recorders (PVRs)

Some set-top-boxes act as personal video recorders capable of recording, time-shifting, and playing back broadcasts. TiVo and ReplayTV are popular stand-alone STB personal video recorders. DirecTV viewers can use Microsoft's UltimateTV set-top-box, which provides PVR functions. The units contain tuners and large hard drives, some capable of holding several days' worth of programming. They are capable of not only recording and playing back, but in time shifting programming. For example, you could set one to record a three-hour game and then start watching the game an hour after it started. As illustrated by stringent copy-protection schemes on videos and DVDs and recent legal action shutting down NAPSTER's streaming audio Web site, a certain stigma associated with digital recording exists. Many in the broadcast industry are fiercely opposed to the notion of PVRs, equating their use as stealing. Similarly, they also oppose features such as ReplayTV's commercial skip function. This hasn't prevented the sales of PVRs, and satellite companies appear to be encouraging their use. Half of the STB/IRDs on DISH Network's Web site contained integrated PVRs (DISH Network, 2002). Perhaps it dawned on satellite companies that it provides a way to sell more of their own programming. While PVR manufacturers are being taken to court, Nielson Media Research has signed an agreement to begin tracking their use.

Spam Television

There are dangers to PVR users. In what was labeled a marketing experiment in the UK, BBC and TiVo programmed PVRs to record the TV show *Dossa and Joe*. Viewers were not able to delete the program for a week and the action generated hundreds of complaints. In the United States, TiVo PVRs automatically recorded promos for a Sheryl Crow album and advertisements for Best Buy. TiVo characterized their action as an innovative way for broadcasters to deliver content to their machines and even promised that viewers could expect more in the future.

IMPACT OF ENHANCED TELEVISION

Impact on Broadcasters

Enhancing TV shows with interactive content is a major undertaking for content producers. Regardless of their feelings for the Internet, broadcast and cable networks along with the larger broadcast stations cannot ignore the necessity of providing Web content corresponding to their programming. Every major network is pursuing some form of two-way enhanced television. Development is expensive and time-consuming. It can take nearly six weeks to develop the interactive content for an episode of *C.S.I.* (Pinsker, 2001). Developers must work closely with the show's producers while ensuring they adhere to the technical constraints of the particular platform they are developing for. Content must be compatible with a wide range of STBs and middleware architecture to reach the largest audience possible. Traditional teams involved in the various stages of production must reorganize and learn new skills in order to create interactive content. Broadcasters developing interactive content using Wink must receive specialized training and certification to use the proprietary development tools. On the whole, broadcasters are in the process of investing billions of dollars into new technologies.

Programming

How is programming affected? Ideally, interactive content is not appended as an afterthought, but carefully integrated into the design of the program. Content creation must be rethought from the ground up and addressed in the early phases of production. Broadcasters are faced with the challenge of developing enhanced content for a wide range of programming including news, sports, and episodic and entertainment programming. TV producers are looking for innovative ways to enhance both new and old programming. At the American Film Institute eTV Workshop, production teams demonstrated enhanced versions of *I Love Lucy*, *Sesame Street*, and *P.O.V*.

According to the BBC, the information needs to be obviously beneficial and relevant. Viewers respond more frequently when alerted that interactive content is available. The information should be in small doses and straightforward (BBC, 2002).

Usage of Enhanced Television

Content providers used to be broken down into broadcast networks, broadcast stations, and cable networks. With the rise of multicasting cable and satellite, this distinction has become somewhat blurred.

Broadcast TV Networks

All of the broadcast networks (ABC, CBS, NBC, PBS, and FOX) provide some degree of enhanced programming

Table 1 FCC Licensed Broadcast Stations—December 2001

UHF	Commercial TV	740
VHF	Commercial TV	576
UHF	Educational TV	254
VHF	Educational TV	125
Total		1,695

From *Broadcast station totals* by FCC (retrieved December 12, 2002, from http://www.fcc.gov/mb/audio/totals/bt011231.html). Copyright 2001 by the Federal Communications Commission. Reprinted with permission.

through either single- or dual-screen applications. It has come to be generally accepted that broadcasters who ignore new distribution opportunities such as enhanced television are likely to lose audience members.

Broadcast TV Stations

According to FCC (2001) as of December 31, 2001, there were 1,695 terrestrial television broadcasters in the United States, consisting of commercial and educational UHF and VHF stations (see Table 1). While enhanced television can take place over existing analog transmission towers through the VBI, digital broadcasting's increased bandwidth opens up a realm of additional data opportunities. As of November 2002, 354 stations were broadcasting digitally in the United States.

Cable TV Networks

Cable networks are uniquely positioned to take advantage of cable's existing infrastructure and, since the debut of Qube, have been at the forefront of exploring interactive programming and applications. CNN, TV Land, The Food Network, Turner Classic Movies, and the USA Network have been early adopters of developing and delivering enhanced programming. The FCC's Ninth Annual Report on Competition in Video Markets, released December 31, 2002 (FCC, 2002), reported that as of June 2002, multichannel video program distributors were reaching nearly 69 million households through cable.

Satellite TV Providers

The FCC's Ninth Annual Report also showed that direct broadcast satellite service grew quicker than cable from nearly 16 million households in June 2001 to approximately 18 million households in June 2002. This figure represents more than 20% of all the multichannel video program distributor subscribers. Major satellite providers include DISH Network and DirecTV. As illustrated by DirecTV's channel lineup, a growing number are offering eTV programming (see Table 2).

Impact on Users

The widespread growth of enhanced and interactive television along with its established acceptance in the UK is a testament to the fact that viewers use and enjoy the new features. Ideally, interactive television provides the best that both television and the Internet offer, combined into an integrated, interactive experience. As demonstrated by WebTV, set-top-boxes can effectively combine the two media and even provide users a way to save money by purchasing only one device.

Table 2 DirecTV's Listing of Stations Offering Enhanced Television

Network	Channel	Availability
Bloomberg Television	353	All day
CBS	Varies	Select programming
CNBC	355	All day
CNN	202	All day
Discovery Channel	278	Select programming
E! Entertainment Television	236	All day
ESPN	206	All day
ESPN2	209	All day
Headline News	204	All day
Lifetime	252	Select programming
Music Choice	802, 811, 814, 816–819, 822–824, 828, 841	All day
NBC	Varies	Select programming
Oxygen	251	Select programming
ShopNBC	370	All day
Showtime East	537	All day
Showtime West	540	All day
TBS Superstation	247	Select programming
TechTV	354	Select programming
TNT	245	Select programming

From *DirecTV Interactive Channel Lineup* (retrieved December 12, 2002, from http://www.directv.com/DTVAPP/imagine/DTVInteractiveChannel.jsp). Copyright by DirecTV. Reprinted with permission.

Personalized TV Experience

One of the promises of eTV is a truly personalized television experience. Using electronic program guides and personal video recorders, viewers can customize the programming they receive, deciding what to watch and when to watch it. With PVRs much of the enhanced content is available for later use, as the information is embedded into a portion of the video signal.

Privacy

While viewers interact with enhanced TV programs, middleware and STB application developers are quietly recording what they watch. For advertisers it provides a wealth of personalized data and a means to deliver targeted, personalized advertising (Chester, Goldman, Larson, & Banisar, 2001). Well within the capabilities of existing eTV system architecture, data profiling systems can record what is being watched and what Web sites are visited along with the time of day and duration. It is chilling to some to consider that marketing agencies can collect personalized data on viewers who are doing nothing more than watching TV.

IMPACT ON E-COMMERCE

Before the dawn of cable, satellite, and the Internet, there was broadcast television. Early business models of broadcasting were simple, centered on advertising revenues. Stations gave programming to the public for free, while charging sponsors for advertising. The greater the number of viewers, the more they could charge the advertisers. Fifty years later, new technologies have lured viewer's eyeballs to other media. Today, traditional broadcasting is just one player in a cacophony of content providers, competing with cable networks, pay-per-view, video-on-demand, superstations, video rentals, satellite distribution, gaming computers, and the Internet, all vying for the viewers' attentions. If viewers could trace the signal path from the source to their TV sets, they would be surprised at the number of unseen entities, including distribution networks, middleware providers, and STB application developers involved in delivering their programming and supporting the interactive exchange of information. These entities are indirectly funded through the viewer's consumer and media purchases, and subscription fees. Sorting out the stakeholders and tallying the comparative rate structures is a complex task, and return on future investment is in good part based of speculation. One thing is clear: the historic broadcast model supported by advertising revenue is woefully ill suited for enhanced television.

Investment in emerging and untested technologies requires an enormous amount of capital and a great deal of faith. Millions of dollars have been invested into software and hardware to support enhanced and interactive television. Stakeholders can be broadly categorized into content creators, broadcasters, distribution networks, and infrastructure developers. Each sector has its own customers, its own means of generating revenue, and concerns over the economic viability of enhanced television.

Business Models

While traditional broadcast and Internet models still have some relevance, the new technologies coupled with current economic conditions have forced investors to take a hard look at the prospects of interactive television. There are four business models that can be applied to interactive television (CRTC, 2002): advertising, pay-per-use, subscription, and t-commerce.

Advertising

While television broadcasters acknowledge that they may lose viewers by exposing their audience to other forms of electronic media, for most, the potential benefits and promise of interactive advertising outweigh any concerns. Similarly, concerns of viewers zapping commercials with personal video recorders are minimized by the numerous possibilities of alternate and contextual ad placement in interactive programming. Banner-type ads can be placed in EPGs, a central and frequently visited location in the viewer's personalized TV experience. Clicking on the ad could draw the viewer to the sponsor's Internet site or to a "walled garden." Walled gardens are a series of Web pages contextually related to the jumping-off point. They let viewers explore additional information and purchase products and services. Another advertising opportunity so readily used by Wink is the on-screen icon, which signals the viewers that additional information is available. Clicking on the appropriate button on the viewer's remote provides additional on-screen information on the product. If there is Internet connectivity the user could visit a dedicated advertising location (DAL) to make purchases or explore an accompanying walled garden.

Pay per Use (PPU)

Video-on-demand and pay-per-view movies and events are well-established examples of the pay-per-use model. Lesser known but gaining attention is the area of pay-per-play. The well-established video rental business provides a ripe market for potential PPU customers. As consumers become accustomed to the ease of purchasing movies and games through their set-top-boxes, satellite and cable distributors are likely to see their PPU revenues soar. Using alternate distribution networks to purchase video-on-demand, games, and other programs could result in a decline in the video rental business from $11 billion in 2000 to $5.9 billion in 2010 (Renaud, 2002).

Pay per Play

PC gamers have been participating in network battles for years, recently joined by X-Box and Playstation users. Digital broadcasters are just catching on that there is a growing market of fun-seeking users who enjoy the interactive features that iTV provides. Similar to VOD and PPV, pay per play lets users buy whenever the impulse strikes. Sky Active's customers can access pay-per-play games with a few clicks on the remote. Users can participate in betting and sports games and play versions of *Space Invaders* and *Who Wants to Be a Millionaire?*. Two Way TV, an iTV developer, has even announced an adult play-per-play application called *Sex Games*.

Subscription

Users of broadband, cable, satellite, and any other distribution service pay periodic subscription fees. In most broadcast distribution systems, the payment interval is monthly, but daily and weekly time periods may also be used. Other subscription fees can include cost of auxiliary services and program packages, along with hardware such as cable modems and decoding equipment. An advantage of the subscription model is that distributors are not directly dependent upon advertising revenue.

T-commerce

Purchasing goods or making financial transactions through interactive television services is known as t-commerce. While viewers can order products, food, and even place bets, middleware and application companies are tracking the traffic. As C. J. Fredrickson, Director of ATVEF, states

> By tracking local consumer "remote click" activity, local operators/affiliates would have the ability to sell information to advertisers and national programmers, such as demographic data based on actual consumer activity. By accumulating this data, local operators would have a new revenue source providing more specific and immediate information on consumer television habits and activity than has ever been available. (Fredrickson, 1999)

Internet Models

Internet TV should lead to greater efficiencies in five areas (Waterman, 2001): (1) reduced delivery costs and capacity constraints, (2) more efficient interactivity, (3) more efficient advertising and sponsorship, (4) more efficient direct pricing and bundling, and (5) lower costs of copying and sharing. While his study focused solely on Internet propagated television, points (2), (3), and (4) should resonate with enhanced television proponents, especially considering that the Internet is quickly becoming an established provider of television content. A key idea behind enhanced TV is the concept of interactivity, providing opportunities for highly targeted television advertising and sponsorship. The Internet can cater to specialized niche markets providing advertisers with extremely focused sponsorship opportunities. Would a marine supply company rather purchase advertising on a cable network seen by millions or a boating webcasting site seen by thousands? Just as cable and satellite distributors can provide bundles of program packages, the Internet offers equally, if not more customizable options.

CONCLUSION

Navigating through the ever-changing and broadening network of interactive programming has been initially confusing for both the industry and consumers. While some U.S. companies have scaled back, most are moving ahead with plans to develop and/or expand their enhanced television services, and overseas the use of interactive TV services is going strong. European communities have been leading the world in delivering interactive television content and advertising. Researchers have projected that by 2005, TVs will surpass PCs in e-commerce revenues (E-Commerce Times, 2000). In November 2002, AOL launched its first interactive TV service on Sky Active in Europe, bringing instant messaging, e-mail, news, weather, and sports to more than 6 million homes. Research data optimistically suggests that enhanced TV will continue to grow and gain acceptability among consumers in the United States, reaching 65 million users by 2006 (Bird, 2002).

Little is understood at this time of the impact that the changeover to digital broadcasting will have on the growth of enhanced television. According to the FCC, by 2006 TV stations should relinquish the analog spectrum, broadcasting only in digital. Considering digital broadcasts as a string of bits, broadcasters must decide how to most effectively use their limited resources and allotted bandwidth. Expanded enhanced features and television content will likely be vying for the same bandwidth required by HDTV. While the industry stumbles with implementing the changeover and consumers balk at the high price of DTV equipment, the low cost and existing infrastructure of enhanced TV could easily fuel acceptance. Considering the widespread acceptance of iTV in the UK and the wealth of opportunities that convergent media offer to advertisers and consumers through interactive content, it appears likely that enhanced TV will make strong inroads in the near future.

GLOSSARY

Advanced Television Enhancement Forum (ATVEF) A not-for-profit coalition establishing industry standard protocols for enhanced programming.

Back channel (a.k.a. return channel) The physical path that the return signal can take from the set-top-box back to the broadcaster through which interactive TV users can transmit information back to the point of origination.

Bandwidth The amount of data that can be sent through a signal path.

DTV (digital TV) A system of transmitting and receiving digital television.

DVB (Digital Video Broadcast Forum) An organization seeking to establish technical standards for digital television and responsible for DVB-RCC (digital video broadcasting return channel through cable) and DVB-RCT (digital video broadcasting return channel through terrestrial).

Integrated receiver/decoder (IRD) Used to decode signals downloaded from satellite receivers.

Interactive program guide (IPG) An electronic display that can be customized to list available and favorite channels, program names, and program descriptions.

Interactive television (iTV) An umbrella term referring to a range of interactive television services.

IP (Internet protocol) A means of sending packets of information over Internet networks.

MPEG-2 (Moving Picture Experts Group) A version of the MPEG video file format used to deliver digital video.

Megahertz (MHz) A measure of frequency equal to one million cycles per second.

Metadata In the digital television context, data about the programming, such as start and stop times, titles, and information on the upcoming show.

Pay per view (PPV) A service available through cable and satellite distributors that allows viewers to purchase programs, such as movies or sporting events.

Personal video recorder (PVR) A device that allows the user to record, store, and playback video such as television broadcasts from a hard drive.

Set-top-box (STB) A device used with televisions allowing users access to enhanced and sometimes proprietary features such as digital channels, video-on-demand, and Internet access.

T-commerce Monetary transactions made over the television.

Vertical blanking interval (VBI) The time it takes for the electron beam to jump from the bottom of a frame to the top of the next. This invisible portion of the television signal can be used to carry data such as teletext, closed captioning, and even HTML documents for set-top-boxes.

Video-on-demand (VOD) A service that uses server technology to allow users to purchase and watch programs of their choice.

CROSS REFERENCES

See *Electronic Commerce and Electronic Business; Video Streaming; Webcasting.*

REFERENCES

ABC's enhanced TV (2002). Retrieved November 15, 2002, from http://heavy.etv.go.com/

Advanced Television Enhancement Forum (2002). Retrieved November 25, 2002, from http://www.atvef.com/

Advanced Television Systems Committee (n.d.). Retrieved November 25, 2002, from http://www.atsc.org/

American Film Institute (n.d.). Retrieved December 20, 2002, from http://www.afi.com

Bird, J. (2002, September). Interactive television. In *Digital television broadcasting: Perspectives on the future* (chap. 4). Retrieved December 2, 2002, from http://www.lemac.com.au/new/DTV/DTV2.html

British Broadcasting Corporation (BBC) (2002). Retrieved November 25, 2002, from http://www.bbc.co.uk/tv/

Canadian Radio–Television and Telecommunications Commission (CRTC) (2002). *Report on interactive television services*. Retrieved October 22, 2002, from http://www.crtc.gc.ca/eng/publications/reports/interactive_tv.htm

Chester, J., Goldman, A., Larson, G. O., & Banisar, D. (2001). *TV that watches you: The prying eyes of interactive television*. Retrieved February 1, 2003, from http://www.democraticmedia.org/privacyreport.pdf

Davis, J. (2000, February 15). *More people surfing Web while watching TV*. Retrieved December 7, 2002, from http://news.com.com/2102-1040-236906.html

Digital Video Broadcasting Forum (DVB) (2002). Retrieved December 21, 2002, from http://www.dvb.org/

DISH Network (2002, November). *Receivers*. Retrieved December 25, 2002, from http://www.dishnetwork.com/

Faria, G. (2002). *DVB-RCT: The missing link for digital terrestrial TV*. Retrieved February 1, 2003, from http://www.broadcastpapers.com/tvtran/HarrisDVBRCTMissingLink01.htm

FCC (2001, December 31). *Broadcast station totals*. Retrieved December 1, 2002, from http://www.fcc.gov/mb/audio/totals/bt011231.html

FCC (2002, December 31). *Ninth annual report on competition in video markets*. Retrieved February 3, 2003, from http://hraunfoss.fcc.gov/edocs_public/attachmatch/FCC-02—338A1.doc

Fredrickson, C. J. (1999, November/December). Seven reasons for Web standard TV transmission. *Video Age International, 19*(7). Retrieved February 25, 2003, from http://www.videoageinternational.com/99nov7Reasons.html

Ha, L. (2002, June 17). Enhanced television strategy models: A study of TV Web sites. *Internet Research: Electronic Applications and Policy, 12*(3), 235–247.

Hawley, S. (2002, September 25). What is middleware and why do you need it? *TelephonyOnline.com*. Retrieved February 1, 2003, from http://telephonyonline.com/ar/telecom_middleware_why_need/index.htm

Hurst, B. S. (2000, July 19). *Add value, not gimmicks, with eTV*. Retrieved May 8, 2003, from I-Marketing News http://www.dmnews.com/cgi-bin/artprevbot.cgi?article_id=9384&mhi=8505&dest=archives

InteractiveTV Today [itvt] (2002). Retrieved January 30, 2003, from http:// www.itvt.com

ITV Marketer (n.d.). Retrieved November 25, 2002, from http://www.itvmarketer.com/

Jones, J. (2002, January 14). *DVB-MHP/Java TV™ data transport mechanisms.* Retrieved May 7, 2003, from http://crpit.com/confpapers/CRPITV10Jones.pdf

Microsoft TV (2002, November). *Enhanced TV*. Retrieved November 28, 2002, from http://www.microsoft.com/tv/Microsoft-TV/Vision/enhanced.asp

National Cable & Telecommunications Association (n.d.). Retrieved January 30, 2003, from http://www.ncta.com

Pinsker, B. (2001, February 22). CBS tinkers with interactive content. Retrieved May 7, 2003, from http://www.thestandard.com/article/display/0,1151,22408,00.html

Renaud, J.-L. (2002) What future for packaged media? Retrieved December 12, 2002, from http://www.advanced-television.com/pages/pagesb/featDVD.html

Rudnick, M. (2000, December 21). *ACTV commences enhanced TV patent infringement suit against Disney*. *DM News*. Retrieved December 4, 2002, from http://www.dmnews.com

Society of Motion Picture and Television Engineers (SMPTE) (2002). Retrieved January 30, 2003, from http://www.smpte.org/

USA Today (2002, April 16) *Interactive TV gets Emmy recognition*. Retrieved May 7, 2003, from http://www.usatoday.com/tech/news/2002/04/16/internet-emmy.htm

Zdnet News (2001, April 29). Interactive TV: Ready, set click. Retrieved December 6, 2002, from http://zdnet.com.com/2001-11-0

FURTHER READING

Swedlow, T. (2001, February 2). 2000: Interactive enhanced television: A historical and critical perspective. Retrieved January 30, 2003, from http://www.itvt.com/etvwhitepaper.html

Waterman, D. (2001). The economics of Internet television. *Info: The Journal of Policy Regulation, and Strategy for Telecommunications, Information, and the Media, 3*(3), 215–229. Also forthcoming in E. Noam, J. Gobels, & D. Gerbarg (Eds.), *Internet Television*. Mahwah, NJ: Lawrence Erlbaum.

Enterprise Resource Planning (ERP)

Zinovy Radovilsky, *California State University, Hayward*

INTRODUCTION TO ERP

Definition of ERP

Few recently introduced software applications have had such a profound impact on business as enterprise resource planning (ERP). ERP systems experienced significant growth in sales in the last decade of the 20th century. From 1990 to 2000, the ERP sales have grown from $1 billion to almost $30 billion. In 2004 the enterprise applications market will reach, according to some forecasts (AMR Research Center—Press Center [AMR], 2000), $78 billion. ERP systems will continue to be one of the largest, fastest-growing, and most influential players in the IT field well into the new millennium.

Enterprise Resource Planning (ERP) is defined as an integrated computer-based system that manages internal and external organization resources. These resources include tangible assets, financial resources, materials, and human resources. At the same time, ERP is an application and software architecture that facilitates information flows between various business functions inside and outside an organization and, as such, is an enterprise-wide information system. Using a centralized database and operating on a common computing platform, ERP consolidates all business operations into a uniform system environment.

The word “enterprise” in the ERP name represents the fact that this system integrates and automates processes within an entire organization regardless of the organization’s nature. In fact, ERP systems have been implemented in manufacturing, distribution, transportation, education, healthcare, banking, and other industries. The word “resource” in the ERP name reflects the system’s intention to rationalize the usage of an organization’s resources. Finally, the word “planning” describes one of the main functions of resource management, i.e., planning resources through a variety of business processes.

Introduced in the early 1990s, the term “ERP” does not reflect the real capabilities of the system it represents. First, ERP systems provide not only planning but also other management functions such as organizing, controlling, scheduling, reporting, and analyzing business processes. Second, a traditional approach to ERP considers it a “back-end” computerized system for managing the internal resources of an organization. However, the ERP has crossed the boundaries of being just a system for planning internal resources. It may often contain “front-end” applications of managing customers and improving customer satisfaction—customer relationship management (CRM); collaborating with suppliers through applications of supply chain management (SCM); and utilizing business-to-business (B2B) e-commerce. As such, this integrated system combines external (front-end) and internal (back-end) business applications and should be defined as an extended enterprise management system or extended ERP system (see Figure 1).

As the original term “ERP” does not fully describe existing integrated solutions, many ERP software vendors avoid using this term to represent their extended enterprise management systems. For example, these systems are called “Total Integrated Solutions,” “E-Business Platform,” “E-Business Suite,” etc. Regardless, the term “ERP” or “ERP system” continues to be the most popular term to describe a suite of integrated applications of an extended enterprise management system.

Brief History of ERP Evolution

ERP systems have more than a 30-year history of evolution. Understanding this evolution is important for comprehending current ERP systems and seeing perspectives of their future development. The idea of integrating and automating business processes utilizing computer programs was first introduced in the late 1960s–early 1970s with the development of *material requirements planning (MRP).* This coincided with the time when manufacturing companies started to extensively employ computers, specifically mainframes and minicomputers, in business and management decisions. MRP is an integrated computer-based system for calculating material and delivery schedules. It combines inventory management, materials planning, capacity planning, purchasing,

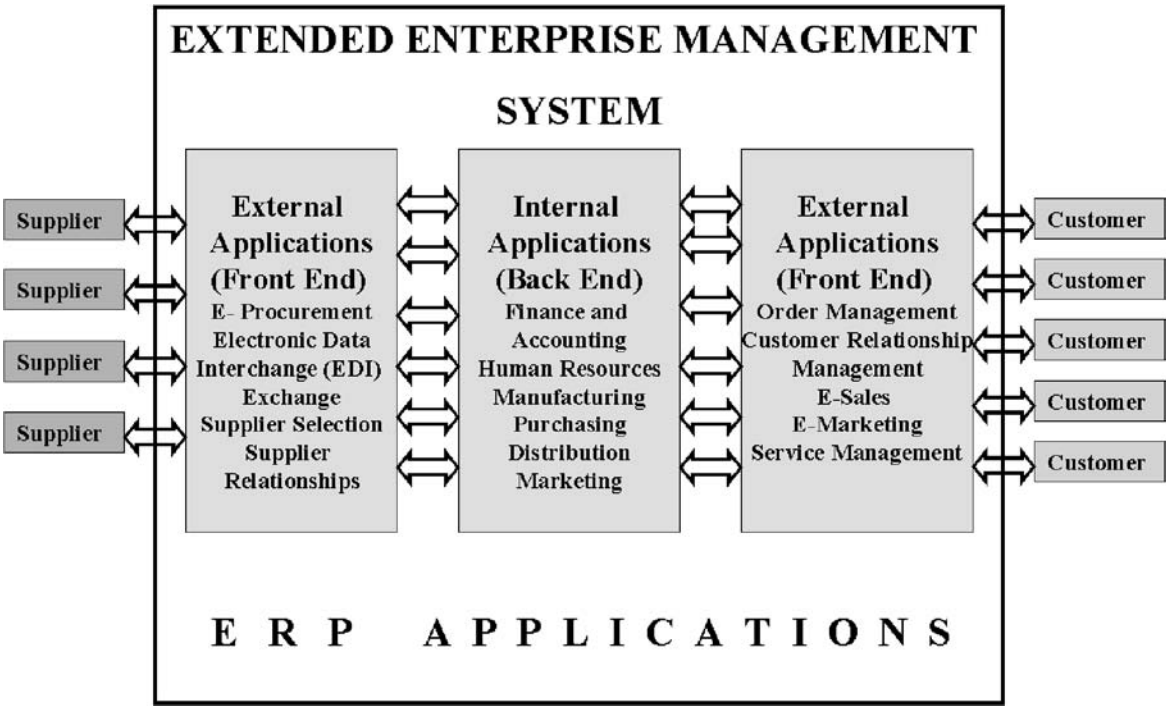


Figure 1: General ERP structure.

and distribution. Later MRP systems also generated purchase orders to suppliers and work schedules for internal production.

The next major step in the evolution of ERP was the development in the early to mid-1980s of a new breed of integrated systems: *manufacturing resource planning (MRP II).* This system fully integrated materials and capacity planning with distribution planning, sale ordering, marketing, finance/accounting, and human resources. Besides planning material and capacity resources, MRP II integrated inventory with financial/accounting transactions, sales orders with materials planning and accounting/finance transactions, marketing and sales analysis with demand forecasting, etc. All inputs and outputs, analysis, and reporting were based on one centralized database system. MRP II was very popular and widely utilized by a variety of large and midsize companies that employed predominately mainframe and minicomputer technology.

Modern ERP systems were established in the early 1990s. This coincided with the proliferation of personal computers, which replaced the old mainframe capabilities with the new client-server technology. Considered as an advanced MRP II system, modern ERP has several major differences from MRPII systems: utilization of the client-server technology and its ability to run on personal computers (clients) and powerful servers utilizing multiple operating platforms (Unix, NT, etc.); advent of ERP into nonmanufacturing companies ("enterprise" rather than "manufacturing" resource planning); and integration of MRP II applications with new business processes like supply chain management and customer relationship management. In the mid- and late 1990s, a large number of companies converted their existing computer systems to ERP due to the Y2K problem. Rather than spending a substantial amount of money on fixing this problem in the existing systems, companies preferred to implement modern ERP systems that were Y2K compliant.

The latest evolution of ERP systems started at the end of the 1990s with the introduction of Internet-enabled ERP systems. This introduction correlated again with major changes in IT systems that were signified by the Internet and e-commerce revolution. The new systems were characterized by Internet-enabled ERP architecture; new front-end e-commerce solutions; easy access to the system by employees, customers, and suppliers; collaborative planning and scheduling; and optimized operations, finance, and marketing decisions.

ERP and E-commerce

The majority of modern ERP systems are fully Internet-enabled systems. This means that communication between a server where an ERP system is installed and many clients (end-user PCs) is done through the Internet.

An ERP system may comprise three main tiers: clients, applications server, and database server (see Figure 2). Clients are end-users that connect to the system via Internet browsers. An applications server incorporates a Web server, forms, system tools, and a variety of ERP programs. A database server includes a relational database with ERP records. Some ERP systems have been developed with separate applications and Web servers, which would define them as four-tier systems (clients, Web server, applications server, and database server).

ERP systems are considered to be a backbone of e-commerce solutions. Successful utilization of the "front-end" e-commerce solutions is unimaginable without

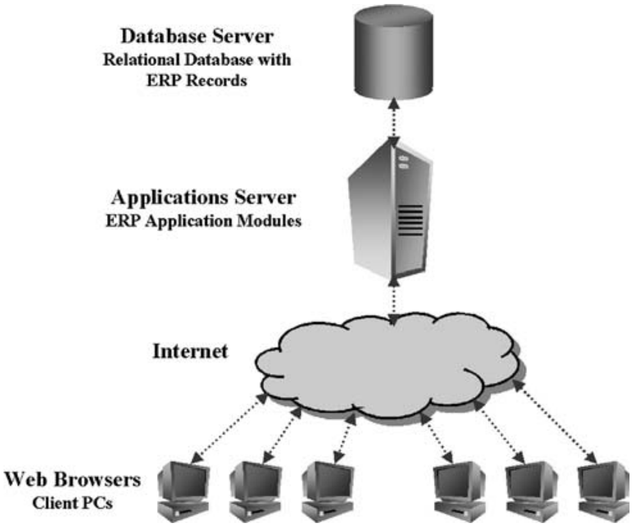


Figure 2: Typical architecture of Internet-enabled ERP systems.

strong support by and cooperation with the internal "back-end" computer systems. In fact, many ERP vendors combine ERP applications and e-commerce solutions in one integrated computer system.

PRINCIPLES, FEATURES, AND APPLICATIONS OF ERP

Main ERP Principles

In many respects, ERP is a result of modern organizations' efforts in designing management information systems. Various processes of an organization have to be linked together so that whenever a change in an external or internal process takes place, the company is able to adjust all other related processes immediately and effectively. ERP systems enable this to happen, not only at the information systems level but also at the applications level, by utilizing certain principles and features. The two main ERP principles are integration and automation.

ERP integration is based upon these items:

- A single database system operating on a common computing platform. All ERP applications would input data and retrieve information from the same database and all employees would have the same point of access to get necessary information
- An integrated set of commonly designed business applications including manufacturing, distribution, marketing, accounting, finance, and human resources. This set consolidates all business processes into a uniform system environment
- Integration between internal company applications and external applications for accessing customers and suppliers.

ERP automation represents the ability of an ERP system to automatically process business transactions and information between different processes and functions inside an organization, as well as between this organization and its customers and suppliers. The elements of ERP automation are as follows:

- Automated business transactions. For example, these can include calculation of production and material schedules, demand forecasts, inventory levels, and production costs.
- Automatic information sharing between numerous organization functions. Data created in one application become available to other related applications. For example, a new employee input made in a human resources module may be available in other applications like purchasing and marketing, which utilize this employee information.
- Automated recording, monitoring, and reporting of the data generated in ERP.

ERP Features

One of the most important ERP features is that it is a *process-driven system*. In contrast to individual and function-driven computer applications in marketing, finance, or operations, an ERP system integrates these functions into a variety of computer-based processes. They represent real business processes that companies apply to managing resources, working with customers and suppliers, etc. In general, an ERP system may include a variety of business processes such as

- order fulfillment
- production planning and scheduling
- capacity planning
- outsourcing materials from suppliers (purchasing)
- shipping products to the customers
- product costing, payments, and receipts
- managing customers
- selecting and managing suppliers.

Some of these processes are interrelated and dependent on one another (see Figure 3). Integrating these business processes into a company's ERP system provides the necessary environment for running the company in real time with all functions being interoperable in the system.

Other features of ERP systems include

- A *relational database* that integrates all data inputs, transactions, and outputs of ERP systems. This can substantially reduce or even eliminate inaccuracy and inconsistency of records that might have existed in separate individual databases. The Oracle database is the most popular relational database used in ERP systems.
- *Company-wide access to information* from a relational database. Transactions that are taking place in each ERP-driven business process may be visible, in principle, to anyone in an organization. In practice, however, the level of an employee's visibility of and access to ERP processes depends on the employee's role (responsibilities) in the company.
- *Multiple simultaneous accesses* to the ERP system by many users and in various locations. This feature may

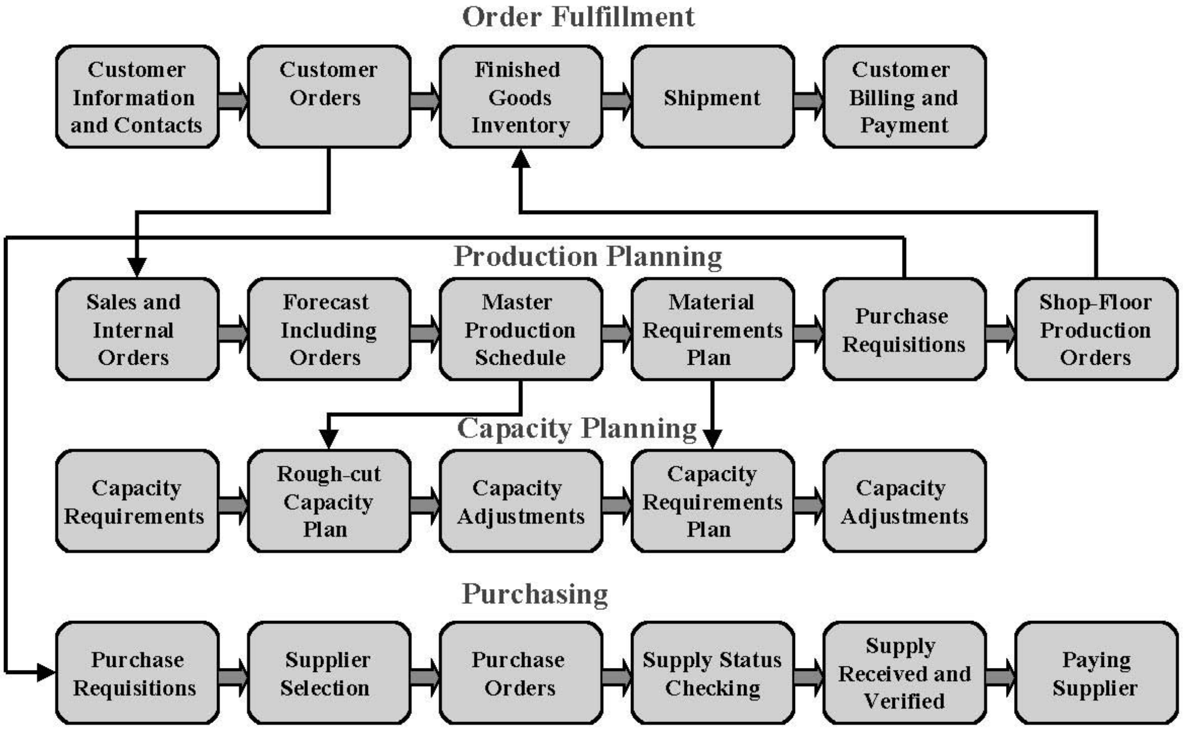


Figure 3: ERP processes and their integration.

be used by a company that has multiple domestic and international locations/divisions and wants to integrate all business functions into one computer system. This feature also enables fast and reliable information sharing between different parts of the company.

- *Scalability*, which means that an ERP system provides adequate capabilities in the situation where the total number of users of the system is growing. This may be due to the company's expansion or to its merger with or acquisition of another company. A scalable ERP system should also accommodate a growing number of users and applications without jeopardizing the speed of transactions or the performance of the entire system.
- *Internet-based architecture* of ERP systems. From a technical standpoint, ERP systems should be flexible enough to run on *various implementation platforms* and operating systems, like Unix, Linux, and NT.

ERP Applications and Their Integration

ERP integrates a great variety of business applications. Their titles, components, and capabilities may vary between different ERP software packages. However, traditionally, an ERP system is composed of *manufacturing and distribution*, *financial, human resource management (HRM),* and *marketing and sales* groups of applications. In addition, an extended ERP system may include *supply chain management (SCM)* and *customer relationship management (CRM)* groups of applications. In many modern systems, the *manufacturing and distribution* applications are tightly associated with, and sometimes become inseparable from, the SCM applications. Thus, they may be considered as one combined ERP group. The same is true for the *marketing, sales, and CRM* group.

Within each group, ERP applications may be clustered into three main categories. The first category—*core applications*—is composed of traditional "back-end" business applications existing in current and old versions of the ERP software. The second category—*applications enhancements*—is a set of additional applications that are used to analyze, improve, and optimize business decisions and processes in each ERP group. The third category—*e-commerce solutions*—is composed of various applications of B2B e-commerce. These groups and categories of ERP applications are presented in Table 1.

The SCM group of applications is associated with planning and scheduling material and information flows from suppliers to a manufacturing/service organization; scheduling, executing, monitoring, and reporting production plans in this organization; and distributing final products to the customers. The applications' enhancements in the SCM group include *manufacturing* and *supply chain intelligence* applications that provide a variety of analyses and data reports on purchasing, supplier scheduling and supplier performance, production results, distribution schedules, and other functions. Besides, this category contains applications of *advanced planning and scheduling (APS)* that enable to create optimized production schedules, taking into account material and capacity constraints. Also, the SCM group integrates a number of B2B e-commerce applications including

- *e-procurement,* which provides capabilities for outsourcing material, labor, and other resources online (online purchasing)

Table 1 Major Groups of ERP Applications

Groups of Applications	Applications' Categories		
	Core Applications	**Applications Enhancements**	**E-commerce Applications**
SCM Group: Manufacturing, Distribution, Supply Chain Management	Inventory Supply Chain Planning Materials Requirements Planning Capacity Requirements Planning Shop Floor Management Warehouse Management Purchasing Quality Management Product Development Management	Manufacturing Intelligence (Analytics) Advanced Planning and Scheduling Supply Chain Intelligence (Analytics) Supplier Relationship Management	E-procurement E-supplier Portals Private Exchange Online Auction Collaborative Commerce (C-commerce) E-design Mobile Supply Chain Commerce
Financials Group: Accounting, Finance	General Ledger Accounts Receivable Accounts Payable Asset Management Cash Management Activity Based Costing Budgeting Property Management Treasury Management	Financial Intelligence (Analytics) Balance Scorecard	E-expenses E-receivables E-payment
HRM Group: Human Resource Management	Personnel Administration Payroll Employee Benefits and Compensations Time Management Training Administration	HR Intelligence (Analytics) Organizational Development	Self-Service Human Resources E-recruitment E-learning HR-centric Web Portal Information Collaboration Mobile Commerce
CRM Group: Marketing, Sales, Customer Relationship Management	Order Management Product Configurator Order Fulfillment Customer Management Service Management Marketing Sales Filed Customer Service	Marketing Intelligence (Analytics) Service Intelligence (Analytics) Advanced Demand Management Customer Interaction Center (Call Center)	E-marketing E-selling E-fulfillment E-service Mobile Sales

- *private exchange,* which is an electronic marketplace that invites suppliers and customers to do trading online
- *collaborative commerce (c-commerce),* which enables buyers and sellers to collaborate on demand forecasting and synchronize plans based on the dynamic exchange of information
- *mobile commerce (m-commerce),* which extends the supply chain management system to all the members of the network with wireless communications regardless of where they are.

Financials are another major group of ERP applications. They include core accounting and financial applications, as well as applications enhancements and e-commerce solutions (see Table 1). Financial intelligence applications are used to analyze and optimize market forecasts, credit risk of new contracts, cash flow and liquidity, financial portfolio, and market risk. Also, financial intelligence applications monitor and track financial performance and measure key financial indicators. The Financials group also contains several B2B e-commerce applications (see Table 1). For example, E-receivables (E-payment) is a secure online application that allows customers to review account information, pay bills, transfer funds, and dispute invoices online.

The HRM group of applications is composed of core applications, modern analytical tools (*HR intelligence/analytics*), and a variety of HRM e-commerce applications (see Table 1). Examples of HRM e-commerce applications are

- *self-service human resources,* which allow line managers and employees to update and use HR information through interfaces personalized to their roles

- *e-recruitment,* which is a set of comprehensive recruitment tools and services that are integrated with a network of recruitment service providers.

The CRM group of applications includes a set of core, enhancements, and e-commerce applications related to customer interactions with the company, managing customer demands, and improving customer satisfaction (see Table 1). Among e-commerce CRM solutions are

- *e-marketing,* which enables implementation and execution of personalized, real-time mass marketing over the Internet
- *e-selling,* which provides comprehensive capabilities for selling products and services over the Internet
- *e-service,* which offers customers and business partners online access to dedicated information such as product catalog and pricing solutions.

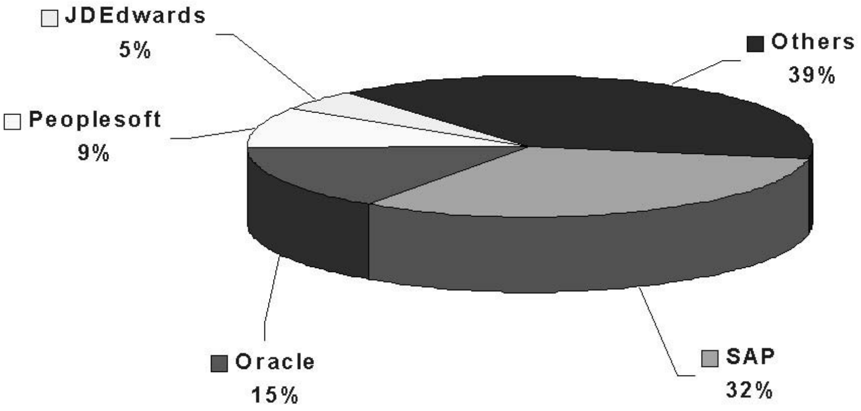
Major ERP Vendors

The number of software companies that develop ERP software runs in the hundreds. Many of these companies provide limited ERP functionality as market niche players; others focus on some vertical markets (industries). However, a relatively large segment of these companies provide comprehensive solutions to their customers. The largest and most popular ERP vendors are SAP, Oracle, and PeopleSoft. These three companies have around 56% of the global ERP market share (see Figure 4).

SAP (www.mysap.com) is the largest ERP vendor in the world (Top 100 Software Vendors, 2001). The most famous SAP system, SAP R/3, includes all common groups of the ERP applications. SAP is specifically popular with large manufacturing companies that run its manufacturing, distribution, and supply-chain applications. Realizing that its R/3 architecture could not meet the challenges of e-commerce in the late 1990s, SAP started implementing an Internet-based solution. For that, SAP launched its totally redesigned, Internet-based solution—mySAP.com. It is a broad undertaking by SAP to extend its back-office suite to the Web. This site is a single entry point for a range of front-end, back-end, and e-commerce ERP applications. The SAP software has been implemented in a variety of companies (predominately, large companies) including IBM, Hewlett–Packard, Philips, Siemens, Chevron, and many others.

Oracle Corporation (http://www.oracle.com) is the second largest ERP vendor. Oracle was the first company to offer a fully integrated, Internet-enabled suite of business applications—E-Business Suite. This system runs entirely on the Internet, supporting e-business across the SCM, Financial, Human Resources, and CRM groups. However, Oracle's Financials and Customer Relationship Management groups of applications are considered to be the company's most popular among the customers. Oracle's E-Business Suite, along with SAP, is also commonly used by large and mid-size manufacturing and service companies. These include Cisco, Sun Microsystems, Applied Materials, Visa International, Alcoa, Xerox, and many others.

PeopleSoft, Inc. (http://www.peoplesoft.com) is also one of the largest ERP vendors. Just like Oracle ERP, the latest version of the PeopleSoft ERP system is on the Internet and represents a full set of Web-based groups of applications. Both Oracle and PeopleSoft put every aspect of a client's business, from supply chain management to sales, on the Internet. However, PeopleSoft considers its ERP system to be the only pure HTML-based solution in the industry. PeopleSoft's applications do not require Java downloads (as Oracle does) to run the browser and are fully accessible from a host of mobile devices such as PDAs and cell phones. However, Oracle claims that Java is typically included in a Web browser and its E-Business Suite is accessible from mobile devices where it makes sense. PeopleSoft customers include a variety of business companies, universities, and government organizations.



Total Revenue, 2000

Source: AMR Research, 2001

Figure 4: Top ERP vendors.

Among the most famous PeopleSoft customers are 3Com Corporation, Sprint, Hewlett–Packard, Analog Devices, and British Telecom.

Besides SAP, Oracle, and PeopleSoft, there are many other quite popular ERP vendors including J.D. Edwards (http://www.jdedwards.com), Invensys Software Systems (http://www.invensys.com), and Siebel Systems (http://www.siebel.com).

ERP BENEFITS, ISSUES, AND APPLICATIONS SERVICE PROVIDERS

ERP Benefits

The rapid growth and proliferation of ERP systems in the past 10 years indicated the great importance of these systems to businesses. In general, the implementation of an ERP system should increase the reliability of a company's information; provide full access to this information at any point of time; automate a variety of tasks and processes in different organizational functions; improve forecasting, planning, scheduling, and reporting; and facilitate external collaboration with customers and suppliers. These improvements may lead to a shorter order-to-market time; lower inventory levels; more efficient and effective management of resources; and, as a result, major cost savings and increased return on investments (ROI). A publication at *CIO Magazine* (The ABCs of ERP, 2002) suggests that the median annual savings from a newly implemented ERP system are approximately $1.6 million.

A survey conducted by Mabert, Soni, and Venkataramanan (2001) showed that 70% of companies that implemented ERP felt that their systems have been successful or very successful. This survey also found that important improvements due to the ERP implementation are associated with (in a descending order of importance)

- quickened information response time and increased interaction across the enterprise
- reduced order cycle
- decreased financial close cycle
- improved interaction with customers
- improved on-time delivery
- improved interaction with suppliers
- lowered inventory levels
- improved cash management
- reduced direct operating costs.

The same survey also recognized that areas benefiting the most from ERP were availability and quality of information and integration of business operations/processes, as well as the three functional areas of inventory management, financial management, and supplier management and procurement.

ERP Issues

Despite the benefits, ERP systems still remain the most complicated, time-consuming, and costly computer-based systems. Implementation time and software cost depend on the company's size, number of implemented applications (modules), and level of planning and preparation of the implementation process.

A survey conducted by Mabert et al. (2001) showed that the ERP implementation time for a large or mid-size company could vary from 23 to 35 months, on the average. The average implementation cost for a large or mid-size company could be around $20 million, and smaller firms may spend up to $12 million (Willis, Willis-Brown, & McMillan, 2001; Mabert et al., 2001). A substantial portion of companies, while implementing ERP, had major cost overruns exceeding the original estimated budget by an average of 60% (Mabert et al., 2001).

Implementation is only one element of the total cost of ownership (TCO) of an ERP system. In general, TCO includes software license fees, hardware cost, installation and implementation fees (consulting, additional personnel, training, testing, etc.), cost of maintenance, and customer support. The average TCO of an ERP system could run up to $85 million, but for some large organizations could reach $400 million (The ABCs of ERP, 2002). According to one study (Cliffe, 1999), more than 65% of executives believed that ERP implementation had at least a moderate probability of hurting their companies. In several cases, documented in the literature sources, a long and inefficient implementation led to substantial drawbacks in the companies' performance, loss of sales and customers, and strained relationships with suppliers. Among numerous companies that experienced costly problems with their ERP implementations were W.L. Gore and Associates (with PeopleSoft) and Whirlpool, Hershey Foods, Allied Waste Industries, and Volkswagen (all with SAP).

Analysis of ERP implementation practices revealed the following major issues that companies need to deal with while implementing an ERP system:

- ERP is a complex and rather rigid system that may require a company to redesign its business processes. The top ERP systems are based on the best business practices, and a company must accept the assumptions inherent in those systems. This pushes the company toward generic processes established in ERP. Thus, the company needs to ensure that adopting generic processes does not sacrifice its production results, customer service, and competitive advantage.
- Successful ERP implementation is nearly impossible without thorough preparation and efficient management of the implementation process. These require strong executive commitment and involvement in the ERP implementation process from start to finish; an empowered implementation team that needs to define objectives, outcomes, and implementation strategy; clearly developed education and training strategies; and full communication of ERP plans to the company's employees.
- ERP systems require virtually a life-changing experience for everyone involved. A company that implements ERP and its employees need to overcome traditional conservatism in utilizing a previously unknown computer system, which has new input forms, output reports, functionality, etc. Thus, extensive and consistent training of

all involved employees becomes a critically important element of a successful ERP implementation.

ERP Solutions for Industries and Small-Size Companies

A top-ranked ERP system, being a time- and money-consuming system, may not be suitable for implementation in small and mid-size companies. Also, a standard ERP system, which is developed with generalized business processes and applications, may not be appropriate for certain industries. To address these issues, the leading ERP vendors provide a host of solutions that include

- customizing ERP systems for individual industries
- developing ERP solutions for small and mid-size companies
- lowering applications costs by selling individual modules
- providing hosting solutions using applications service providers (ASP).

For example, SAP established customized solutions for 21 different industries, among which are aerospace and defense, automotive, banking, chemical, consumer products, engineering and construction, financial services, and healthcare. Each industry-related ERP system contains industry-specific applications, for example, *profit management* and *risk management* applications for banks or *healthcare e-business applications* for the healthcare industry. Besides, an industry-related system may also include standardized ERP applications like accounting/finance, CRM, human resources, and others relevant to that industry.

Oracle has introduced, along with its well-known E-Business Suite, an integrated set of applications for small- and mid-size companies—Small Business Suite. This system consists of only a few ERP applications limited in capabilities, like *accounting, sales force automation, customer support management, employee expanses, time and billing, payroll, online bill payment, Web store,* and *customer care*. Due to the relatively small number of applications and their limited capabilities, the TCO of the Small Business Suite is lower than that of a general ERP system.

Another way of lowering TCO is to acquire a group of applications rather than the entire system. Analysis of ERP implementation showed that companies tend to utilize only a subset of the applications available in a standard ERP system. A survey conducted by Mabert, Soni, and Venkataramanan (2000) showed that the most frequently implemented ERP applications are financial accounting, materials management, production planning, order entry, purchasing, financial control, and distribution/logistics. Thus, selling these modules individually or in an integrated package may reduce customer TCO and shorten the implementation cycle.

Applications Service Providers (ASPs)

Due to the complexity and costs of ERPs, ERP hosting/outsourcing became a popular option for small and mid-size companies that lack the capital and resources to do in-house implementation. An applications service provider (ASP) is an organization that hosts, manages, maintains, and monitors computer applications on behalf of its customers. An ASP assumes responsibility for the underlying delivery mechanisms, which include networking infrastructure and hardware requirements. An ASP may also be responsible for application maintenance and upgrades, training, technical support, and overall systems management. Typically, an ASP charges an end-user organization a fixed monthly fee based on the application usage and services rendered. These services may include hardware installation, customer support, maintenance, and upgrade.

Many companies outsourcing ERP-related applications evidenced benefits from ASPs. Research of the Aberdeen Group (2001) suggested that hosted solutions could be implemented 23% faster and at 60% lower cost than in-house ERP systems. Research conducted by the author identified the following important benefits of outsourcing enterprise applications through an ASP versus implementing them in-house (the benefits are presented in descending order of their importance):

- *Various cost savings/cost reductions, which improve companies' return on investments (ROI)*. These reduced costs include costs of additional IT staff to manage implementation, maintenance, and support of the applications; expenses of creating, running, and maintaining a complex infrastructure; cost of installing and running servers and other hardware; and cost of software upgrades/updates.
- *Companies can concentrate on their core competencies.* Outsourcing complex ERP systems allows companies to focus on their core strategic activities and achieve greater competitive differentiation in their principal business areas.
- *Fast implementation schedule.* According to various researches, the average time to implement outsourcing solutions is around 6 months, which is at least four to six times lower than the average time for implementing an ERP system in-house.
- *Investment risk reduction.* The ASP model reduces companies' risk of making huge investments in the software implementation and purchasing soon-to-be obsolete software.
- *Regular software upgrades/updates.* In many cases, ASP companies are responsible for the timely upgrade of their software according to customer suggestions and requests.

While selecting an ASP company, customers should consider a combination of factors representing major aspects of the ASP model. Based upon several research reports and information from the ASP-related Web sites, the author identified important factors that customers need to look into while selecting the ASP. These factors are presented below in descending order of their importance to customers:

- *Guaranteed delivery (reliability of vendor).* Customers would need to get a reliable ASP that provided service

Table 2 Main Steps and Elements of ERP Implementation

Steps	Elements
Phase 1. Preparation and Selection of ERP	- Form ERP selection and implementation team - Develop ERP vision (needs, objective, outcomes) - Identify the implementation model (one-vendor solution, best-of-breed solution, outsourcing, etc.) - Develop selection criteria - Establish ERP software candidate list (four to six candidates) - Create Request for Proposal (RFP), and send it to the prospective candidates - Review their responses and identify three-four finalists - Request the finalists to demonstrate their packages - Select the winner and negotiate the contract - Justify the investment
Phase 2. Installation and Implementation	- Organize the implementation project - Define the performance measures for the new system - Create the initial detailed implementation project plan - Educate the project team - Assess integrity of the existing database - Install new or upgrade existing hardware - Install the software; perform the computer room pilot - Educate the ERP users - Define and refine procedures for the new system - Ensure integrity and accuracy of the data - Bring the fist module/product/plant live; refine and adjust. Repeat the same for other modules/products/plants - Improve continually

virtually all the time, with uptime of at least 99% and guaranteed access 24 hours a day and 7 days a week.

- *Quick implementation time*. The implementation cycle time should be no more than 6–9 months.
- *Vendor's reputation*. Customers should select a well-referenced vendor that has a successful track record of implementations, excellent customer service, and maintenance.
- *Financial stability of ASP*. An ASP's financial position is very critical to supporting the ongoing business so that it can provide services over the long term. Therefore, a customer, prior to making its final ASP decision, needs to fully investigate the ASP's financial performance and results.
- *Scalability*. ASPs should offer services that scale as quickly as a company develops.
- *Cost of ASP*. Customers should look for low-cost solutions with a fixed fee structure at compatible rates.
- *Appropriate security*. Customers would need to get ASPs that provide layered data security and backup service. Moreover, only authorized escorted personnel should have access to the data center.
- *Expertise with applications*. ASPs must have extended experience with and expertise in implementing and maintaining applications it hosts.
- *Professional services*. It is preferable to have an ASP company that is a niche player and focuses on a specific market segment. ASPs should have an ability to integrate with existing systems and provide some application customization.
- *Regular updates*. Customers would need to select an ASP that frequently enhances its software with regular updates.

ERP SELECTION AND IMPLEMENTATION

The selection and implementation of an ERP system consists of several major steps that prepare a framework for successful implementation, identify an ERP software vendor and applications to be implemented, and install and implement the software. As described by Langenwalter (2000), the main steps and their elements are presented in Table 2.

One of the critical elements in the first step is to develop a company's ERP vision. The vision should include objectives, needs, and expectations for implementing such a system, as well as an implementation model. Today, companies utilize several implementation models. One of them is a single integrated ERP system from an ERP software vendor, i.e., a "one-vendor" system. Another model is associated with a combination of the "best of breed" applications. For example, a company might decide to select SCM applications from SAP, HR from PeopleSoft, and Financials from Oracle. In this case, separate subsystem solutions are interfaced and not integrated (Harrelson, 2001). The third model might represent an outsourcing of ERP solutions through ASP. To choose a specific implementation model, a company needs to take into consideration a variety of factors including size of the company, industry the company belongs to,

Table 3 ERP Implementation Measurements

Group of Measurements	Possible Indicators
Financial results	- Cost savings - Return on investments - Financial payback - Total cost of ownership
Process improvement results	- Order fulfillment time - Procure-to-pay time - Inventory level - Production/service quality
Customers and supplier satisfaction	- Simultaneous number of customers in the system - Number of customer complaints - Relationship with suppliers
Systems results	- Availability of information - Reliability of information - System response time

ability to integrate various applications, and projected investments.

Selection of the best ERP vendor should be based on the following criteria:

- *Features and functionality*—the capabilities that ERP software has to fulfill the company's needs; the multiplicity of features and applications that the vendor can provide
- *Estimated total cost of ownership (TCO)* as defined in ERP Issues
- *Provided service and support*—quality and timeliness of a vendor's product installation, service availability, customer service, business consulting, and system integration
- *Financial viability* of the vendor to provide long-term service for the company (vendor's revenues, growth margins, earnings per share, fundamentals ratios, etc.)
- *Technical execution*—ability of the vendor to meet industry milestones (R&D capabilities in comparison to those of other vendors).

To monitor and analyze implementation processes and results, a company needs to put in place a system of ERP implementation measurements. This system should describe financial results of implementing ERP, process and system improvements, customer and supplier satisfaction, and technical/system performance results. Possible measurements of ERP implementation are presented in Table 3.

CONCLUSION: THE FUTURE OF ERP

In more than 30 years of evolution, ERP systems have undertaken a significant transformation from function-specific applications to fully integrated process-driven systems. These systems operate over the Internet and contain both back-end and front-end capabilities. Being one of the main enterprise-based computer architecture and business application, ERP systems have been experiencing a dramatic growth in sales for the past 10–12 years, and, according to various forecasts (AMR, 2000; Data Analysis Group, 2000), will continue to grow in the future.

The further development of ERP systems is associated with the following trends:

- Full integration of e-commerce solutions into ERP systems
- Substantial increase in the number of collaborative ERP tools and applications
- Future development of analytical and business intelligence tools imbedded into ERP
- An open ERP architecture environment that makes it possible to connect the system with other internal and external customer and supplier applications.

In lieu of these trends, The Gartner Group (Genovese, Bond, Zrimsek, & Frey, 2001) has introduced and described a new ERP paradigm shift—ERP II. According to Gartner, "ERP II is an application and deployment strategy that expands out from ERP functions to achieve integration of an enterprise's key domain-specific, internal and external collaborative, operational and financial processes." A cornerstone of ERP II is open architecture of its components. That means that the monolithic systems of the past will have to change. ERP II will be more componentized, and, thus, companies will open their systems up to other companies, suppliers, and customers. This would further irradiate the "inward looking" nature of core ERP systems and the "one vendor owns everything" ERP model. According to the ERP II concept, core ERP systems have to be redefined and extended to embrace the Internet, new virtual supply chain models, CRM systems, and the B2B e-commerce models. With ERP II, the role of ERP systems expands from an attempt to optimize enterprise resources to a focus on "exposing the information involving those resources to other enterprises within a community of interest."

The Gartner Group's research (Genovese et al., 2001) predicts that by 2005 40% of ERP II vendors will offer native integration of between 60% and 80% of enterprise-centric commerce processes, depending on domain. Also, by 2005 less than 50% of enterprises will rely on a single vendor to enable more than 80% of their enterprise-centric commerce processes.

GLOSSARY

Applications service provider (ASP) Organization that hosts, manages, maintains, and monitors computer applications on behalf of its customers.

B2B (business-to-business) e-commerce Use of computer networks, primarily the Internet, to buy and sell products, services, information, and communications between business organizations.

Customer relationship management (CRM) Set of applications related to customer interactions with a company, managing customers' demands and orders, and improving customer satisfaction.

Enterprise resource planning (ERP) Integrated computer-based system that manages internal and external organization's resources; application and software architecture that facilitates information flows between various business functions inside and outside an organization.

ERP II Application and deployment strategy that expands out from ERP functions to achieve integration of an enterprise's key domain-specific, internal and external collaborative, operational, and financial processes.

Manufacturing resource planning (MRP II) Integrated computer-based system that combines materials and capacity planning with distribution planning, sale ordering, marketing, finance/accounting, and human resources.

Material requirements planning (MRP) Integrated computer-based system for calculating material and delivery schedules; combines inventory management, materials planning, capacity planning, purchasing, and distribution.

Process-driven system ERP feature of integrating business functions into a variety of processes, which are computerized versions of real business processes that companies apply to managing resources, working with customers and suppliers, etc.

Scalability ERP feature of providing adequate capabilities when the total number of users of a system is growing due to a company's expansion or its merger with or acquisition of another company.

Supply chain management (SCM) System of principles, methods, and applications of planning and scheduling material and information flows from suppliers to a manufacturing/service organization; scheduling, executing, monitoring, and reporting production plans in this organization; and distributing final products to the customers.

Total cost of ownership (TCO) Total cost of owning a computer-based system; includes software license fees, hardware cost, installation and implementation fees (consulting, additional personnel, training, testing, etc.), cost of maintenance, and customer support.

Y2K problem Derived from presenting a year as a two-digit number in software; for example "00" could stand for "1900" or "2000." This can cause failures in arithmetic, comparisons, sorting, and input/output to databases or files when date-related data are manipulated. This problem potentially exists in old computer-based systems.

CROSS REFERENCES

See *Application Service Providers (ASPs); Business-to-Business (B2B) Electronic Commerce; Customer Relationship Management on the Web; Supply Chain Management.*

REFERENCES

AMR Research Center—Press Center (2000). *AMR research predicts enterprise applications market will reach $78 billion by 2004.* Retrieved 2000 from http://www.amrresearch.com/pressroom/files/00613.asp

Aberdeen Group (2001). *E-procurement: Finally ready for prime time* (Aberdeen Group Viewpoint, March 21).

Cliffe, S. (1999). ERP implementation. *Harvard Business Review,* (1), 16–17.

Data Analysis Group (2000). *Forecast enterprise resource planning (ERP) sales to grow at 24.9% to $73 bln in 2004.* Retrieved October 24, 2000 from http://www.infotechtrends.com/prenterpriseresourceplanning.htm

Genovese, Y., Bond, B., Zrimsek, B., and Frey, N. (2001). *The transition to ERP II: Meeting the challenges* (Strategic Analysis Report, September 21, 1–34). Stamford, CT: Gartner Group.

Harrelson, B. (2001, June). Decisions: Comparing best-of-breed solutions to all-in-one ERP suites. *APICS-The Performance Advantage,* 25–27.

Langenwalter, G. A. (2000). *Enterprise resource planning and beyond: Integrating your entire organization.* Boca Raton, FL: St. Lucie Press.

Mabert, V. A., Soni, A., and Venkataramanan, M. A. (2000, second quarter). Enterprise resource planning survey of U.S. manufacturing firms. *Production and Inventory Management Journal,* 52–58.

Mabert, V. A., Soni, A., and Venkataramanan, M. A. (2001, third/fourth quarters). Enterprise resource planning: Measuring value survey of U.S. manufacturing firms. *Production and Inventory Management Journal,* 46–51.

The ABCs of ERP (2002). *CIO Magazine.* Retrieved February 7, 2002 from http://www.cio.com/forums/erp/edit/122299_erp.html

Top 100 Software Vendors (2001). Retrieved 2001 from http://www.manufacturingsystems.com/top100/

Willis, T. H., Willis-Brown, A. H., and McMillan, A. (2001, second quarter). Cost containment strategies for ERP system implementations. *Production and Inventory Management Journal,* 36–41.

E-systems for the Support of Manufacturing Operations

Robert H. Lowson, *University of East Anglia, United Kingdom*

INTRODUCTION

In five years, the "e" will disappear from the term e-manufacturing as there will no longer be anything new or mystical about "e-manufacturing." However, despite being increasingly evident in management literature, the subject has received little clarification concerning its full meaning, application, potential and challenges—something this chapter hopes to rectify.

This chapter offers five distinct, yet interconnected themes. First, an in-depth analysis of e-manufacturing covers its background, the concept, the potential, and the implications for manufacturing of being able to produce at disparate sites, often globally. Linked to this, the second part evaluates the strategic choices possible when using a global production strategy. The third section further develops the strategic argument by assessing how an e-manufacturing strategy can be formulated. A number of steps are considered and a direct link is then formed to the penultimate section: competitive advantage from e-manufacturing. Finally, empirical evidence for the application of e-manufacturing is given in the form of a small case study.

THE CONTRIBUTION OF E-MANUFACTURING

Before turning to the contribution of e-manufacturing, a word of caution is necessary. The Internet is an extremely important new technology that has received much attention through the latter part of the 20th and the early 21st centuries. However, as Porter (2001) points out, it is perhaps time to take stock and develop a clearer view. As he suggests: to move away from the rhetoric about "Internet industries" and "e-business strategies," and a "new economy." To see the Internet for what it is: "an enabling technology—a powerful set of tools that can be used wisely or unwisely in any industry and as part of almost any strategy." However, technology and the Internet, in particular, do provide a better opportunity to establish operational effectiveness than previous generations of information technologies.

The Internet per se will rarely provide a competitive advantage. The creation of true economic value (the gap between price and cost to produce) will always be the bottom line in terms of survival or failure. Sustainable competitive advantage can only be achieved by operating at lower cost, by commanding a premium price, or doing both. These cost and price advantages can be achieved in two ways: First, *operational effectiveness* (doing the same things as your competitors but doing them better); and second, *strategic positioning* (doing things differently from competitors in a way that delivers a unique type of value to customers). Best practice in terms of operational effectiveness includes better technologies, superior inputs, better-trained employees, more effective management structure, and a clearly articulated operations strategy that links business policy to operational activity.

The Internet can affect operational effectiveness. A powerful tool, but, according to Porter, "simply improving operational effectiveness does not provide competitive advantage . . . this can only be done by achieving and sustaining higher levels of operational effectiveness than competitors." Best practice tends to be copied quickly! As it becomes harder to sustain operational advantages, strategic positioning becomes all the more important. This goes far beyond the pursuit of best practice (the quest of the operations strategy). Strategic positioning involves the highly integrated configuration of a tailored value chain, the series of primary activities required to produce and deliver a product or service: inbound logistics, operations, outbound logistics, marketing and sales, and after-sales services—the first three also being the province of the operations strategy.

According to this viewpoint, the Internet is ostensibly a powerful new technology to aid operational effectiveness, but in reality, is just another way of doing business and not a business strategy in itself. It is,

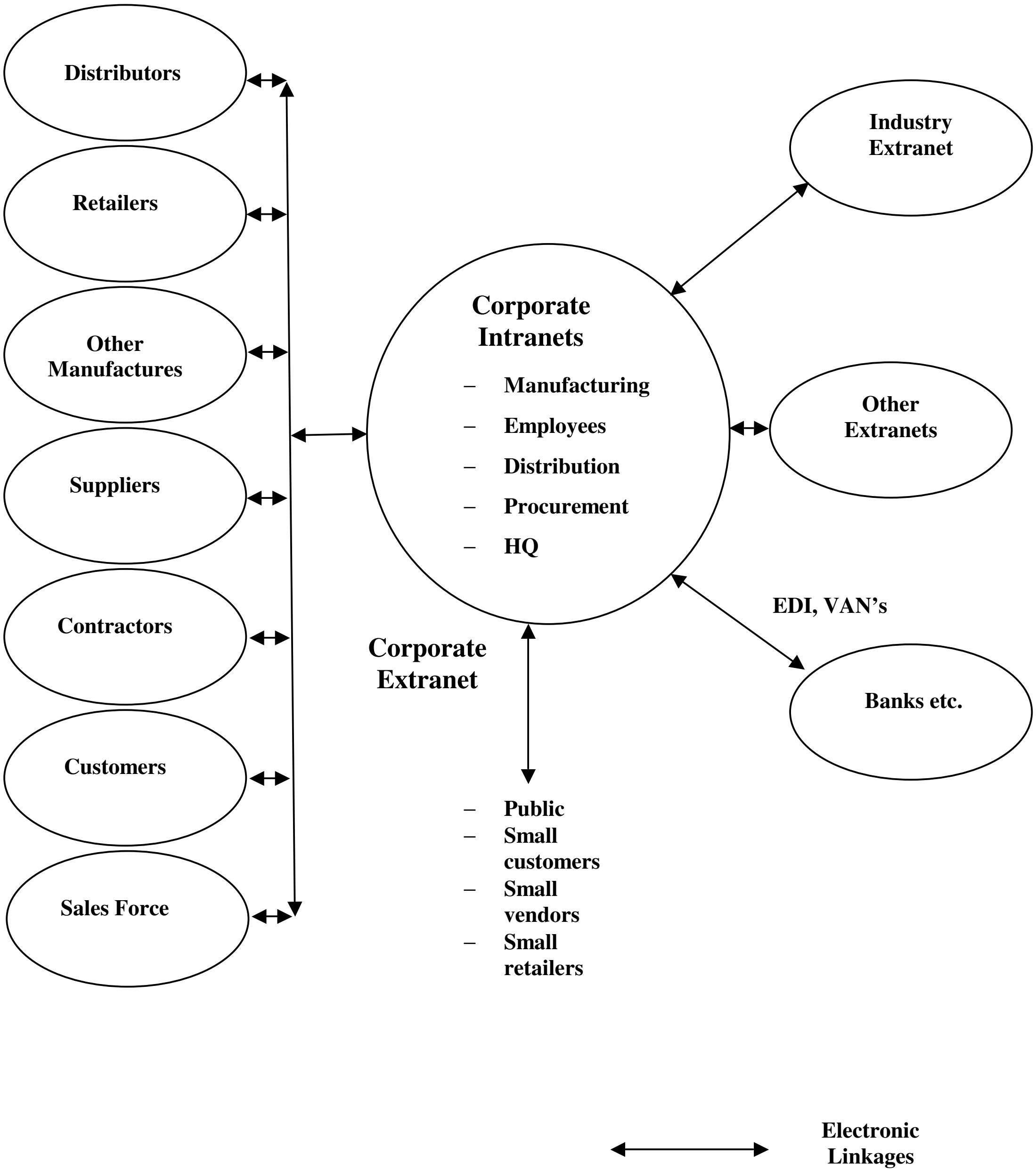


Figure 1: The networked model.

nevertheless, an operational management resource that can be advantageously exploited as part of an operations strategy.

The Background to E-manufacturing

Electronic commerce can change mass production systems to demand-driven and customized, just-in-time manufacturing. Furthermore, production can now be integrated with finance, marketing, and other functional systems, as well as with business partners and customers. Using Web-based enterprise resource planning (ERP) systems, orders are taken from customers and directed to designers (using computer-aided design) and/or to the production floor, within seconds. Production cycle times can be cut by 50% or more in many cases, especially when manufacturing is completed in a country different from the design and engineering location. Companies like IBM, General Motors, General Electric, and Boeing are assembling products for which the components are manufactured in many locations. Sub-assemblers gather materials and parts from their vendors, and may use one or more tiers of supply manufacturers. A whole disparate production system is capable of being electronically coordinated from one central location. Using electronic bidding, assemblers acquire subassemblies 15–20% cheaper and 80% faster. In addition, such systems are flexible and adaptable, allowing fast changes with minimum cost. Further, costly inventories that are part of a mass-production era can be minimized. As can be see in Figure 1, firms today can combine the Internet, intranets, and extranets in an integrated whole. The manufacturing enterprise in this figure is part of a number of extranets. Apart from its own, there may be connections to banks and other financial institutions using secured electronic data interchange (EDI) across a value- added network (VAN). VAN-based EDI systems can also link the organization with its suppliers and other business partners; meanwhile, Internet-based EDI is used with smaller firms.

E-manufacturing is highly dynamic in nature, but will typically include certain critical systems, technologies, and software, e.g., computer-aided design, manufacture, and engineering; remote failure effects analysis; manufacturing resource planning and enterprise resource planning; just-in-time; outsourcing; vendor-managed inventory, and e-supply chains.

There are numerous interpretations of the e-manufacturing concept, some more specific than others. E-manufacturing is the application of open, flexible, reconfigurable communications and computing technologies to manufacturing practices and the movement and storage of products. In its highest form, it is a seamless integration of e-business components that can coordinate distributed manufacturing operations as a single entity, matching demand to supply—"cybermanufacturing" (Toussaint & Cheng, 2002). It should be noted that there are differences between e-manufacturing, e-supply chain, and e-logistics. The following perspectives are adopted in this chapter:

E-manufacturing—Electronically mediated data/information exchanges and the information communication technology systems that support them. To facilitate the operation, enhancement, reengineering, and integration of the range of production processes, both within an organization and externally (both supply and demand sides), takes place. Physical manufacturing is local, regional, or international with the global sharing of knowledge using the Internet. E-supply chains and e-logistics are the interfaces to manufacturing.

E-supply chain—Supply chain management, part of a wider supply network, that coordinates the flow of materials both up- and downstream. Equally important are the information streams involved. Once these flows are enhanced by an electronic means of communication, then the e-supply chain is evident.

E-logistics—Logistics deals with microissues (such as transportation) that take place after production; in other words, the demand side. Information flows are vital and if electronic means are provided to support them, e-logistics becomes available.

The remote monitoring and maintenance of a production process seems to have instigated the birth of e-manufacturing. Today, adherents suggest (Pires, 2001) that it encompasses a new business model. Web-based customer order systems activate procurement and manufacture processes by digitizing the raw data and information associated with the production life cycle; these include computer-aided design, digital quality control, remote maintenance, and real-time shop floor information and control systems.

The E-manufacturing Concept

E-manufacturing can have two applications: tactical and strategic. In the former, the Internet is used to monitor processes on the manufacturing floor and other material handling to ensure optimal operation. This monitoring can be undertaken on a remote basis, providing basic information regarding work in process, inventory, and other problem diagnostics. At a strategic level, however, the impact of e-manufacturing is likely to be more profound. Here, e-based systems across a supply network provide electronically transmitted knowledge to design products, transmit orders and enable order management, procure components and materials, link planning and scheduling, and allow for the testing and validation of finished products. It is business process visibility of capacity, processing, and inventories, from raw material to finished goods, and the collaboration involved, that is likely to bring the greatest strategic benefits. Yet, the transformation from full internal integration to supply chain and customer integration is far from easy. Merely overlaying Internet technology upon existing processes will not suffice.

In a fully implemented e-manufacturing scenario, the strategic benefits offer real opportunities for competitive advantage in linking disparate, world-wide production sites and supply networks in a coordinated move closer to mass customization and the individualization of goods; see Figure 2. In this figure, e-manufacturing can be visualized as a hub activity. Using electronic data/information flows, the e-business structure is linked to integrated product development and moves closer to mass customization (including intelligent products and technologies). It seems clear that e-manufacturing has immense potential for opening a number of new markets.

E-manufacturing Market Potential

> E-commerce has been mismanaged at the manufacturing level. People have figured out all kinds of ways for customers to order instantly. But they haven't figured out a way to actually know in advance, on a real time basis, whether or not they will be able to satisfy those orders. (T. Murrell, personal communication, July 15, 2002)

It is conceivable that the emergence of factory-to-business (F2B), business-to-business (B2B), and factory-to-consumer (F2C) models may alleviate the dire problems described in the quote above in providing the missing "e-fulfillment link," as seen in Figure 3. The figure show the conceptual position of a firm locating its physical operations globally and serving its customers (using F2B, B2B, and F2C) using e-manufacturing. Other linkage mechanisms between product and the customer will be found in the appropriate operations strategy such as, for example, supply chain management (SCM), vendor-managed inventory (VMI), and efficient consumer response (ECR) (Lowson, 2002a). These strategies are themselves enhanced by various internet applications.

Manufacturing and Production for E-business

It is now possible to examine the nature of manufacturing in an e-business environment. The process can be defined

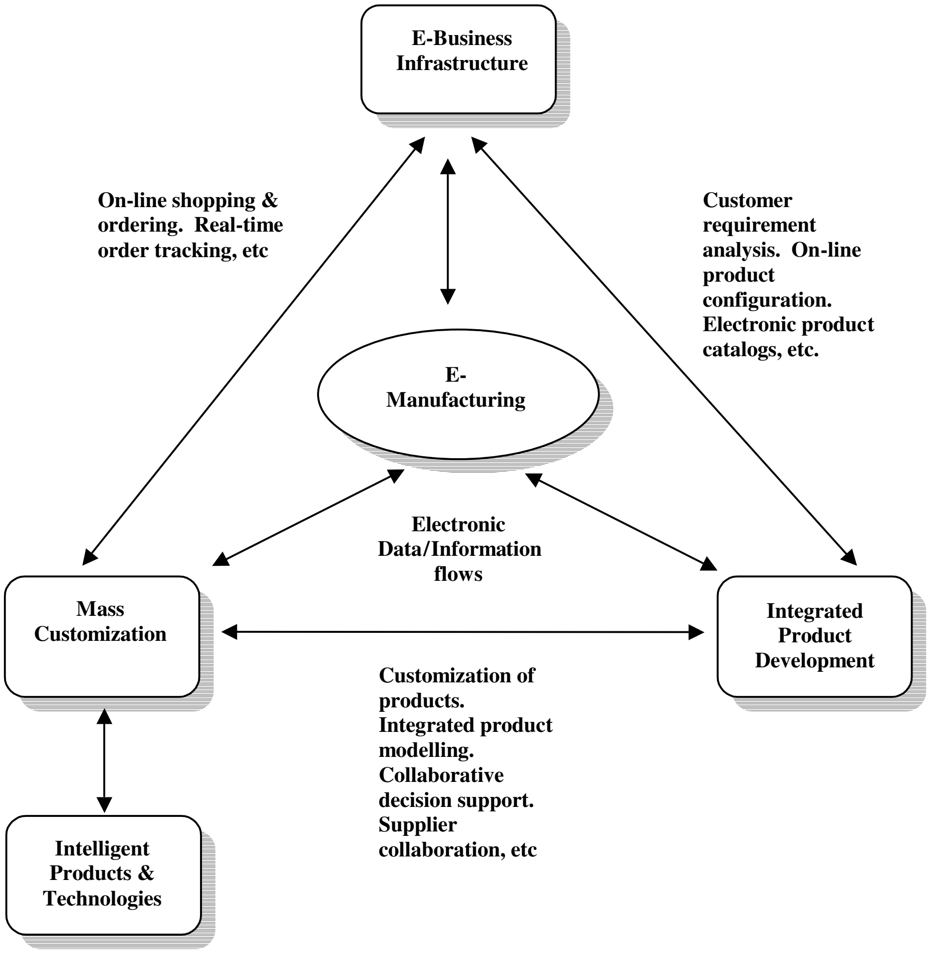

E-Business Infrastructure
On-line shopping & ordering. Real-time order tracking, etc
Customer requirement analysis. On-line product configuration. Electronic product catalogs, etc.
E-Manufacturing
Electronic Data/Information flows
Mass Customization
Integrated Product Development
Customization of products. Integrated product modelling. Collaborative decision support. Supplier collaboration, etc
Intelligent Products & Technologies

Figure 2: The e-manufacturing model.

as follows: “The management of the internal and external systems, resources and technologies that create and deliver the firm’s primary products.”

This definition expands the manufacturing concept beyond just internal processes. It can encompass other activities such as purchasing, distribution, and product and process design. Further, in an e-business setting, there will also be external managerial responsibilities at a supply network level, covering the interconnections between external firms.

E-business can revolutionize many elements of production management due to its effective time, resource, and cost reduction. This is mainly achieved through improved communication and dissemination of economically valuable data and information by improving visibility and integration of supply chains and effectively replacing the need for inventory with real-time information (Van Hoek, 2001).

Such improvements provide opportunities for e-manufacturing to further increase economic efficiencies by matching output to demand and facilitating the exchange of information and goods and services—a move closer to the perfect market.

E-manufacturing Strategy

In order to understand the organizational context of e-manufacturing and an e-manufacturing strategy, the infrastructure of the modern organization is viewed as consisting of four conceptual levels:

1. *Business model*—Definition of a general, long-term business strategy;
2. *Operations strategy*—Decisions made regarding the medium- to long-term operational or manufacturing aspects of providing certain products and services;
3. *Operational management*—The tactical processes necessary to implement both the above strategies and provide products and services; and
4. *Information system (IS) and information technology (IT) architecture*—The supporting information and technology necessary in an e-business setting (often there might also be distinct IS and IT strategies).

The role of the operations strategy and operational management (levels 2 and 3) is to execute or implement the general business strategy and effectively use the tools and information flows involved in level 4. An e-business must also seek to understand how to take advantage of new operational and IS/IT capabilities to migrate to more effective business models. The challenge for an e-manufacturing strategy is to translate these business models into activity.

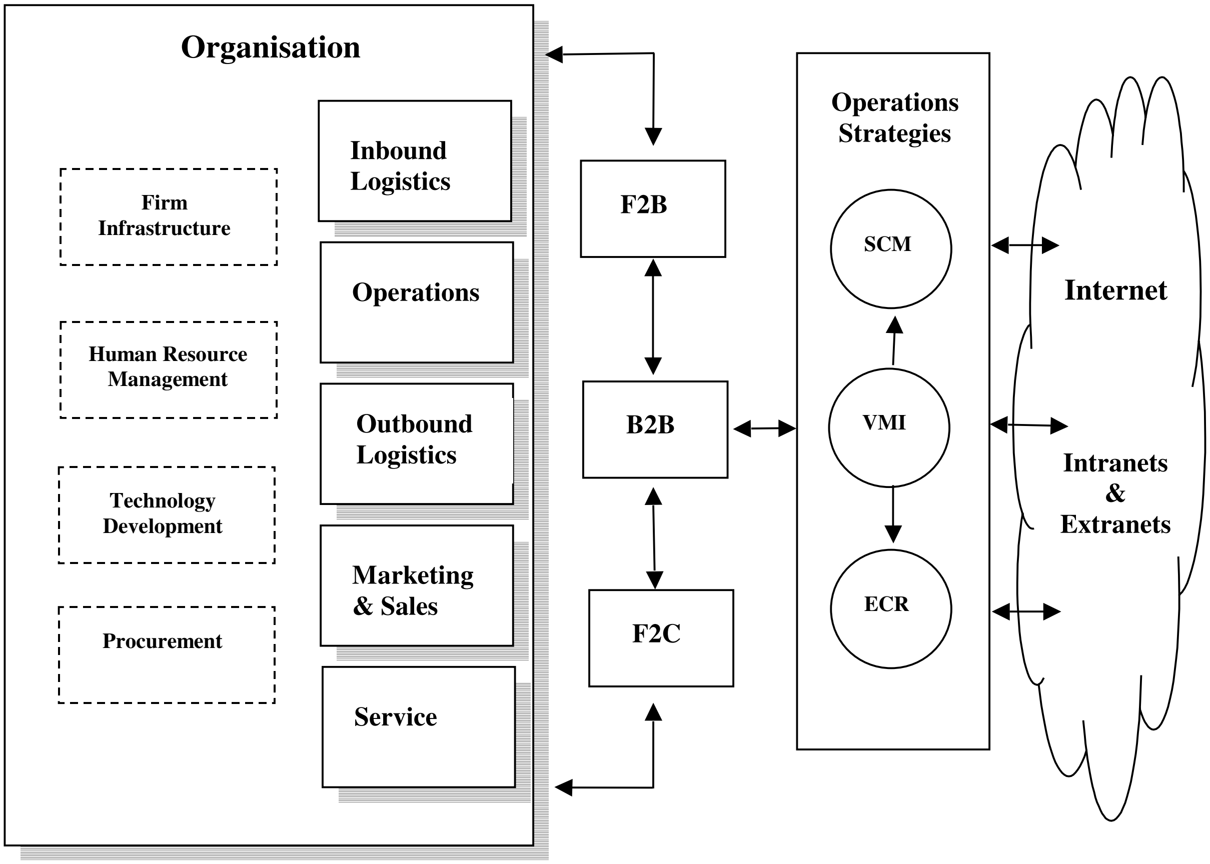


Figure 3: F2B, B2B, and F2C: A 3-tier model.

An operations strategy (in this case one for e-manufacturing) is concerned with both the internal and external decisions necessary to translate the business vision. The subsequent business model will be a fluid collection of business activities—each one focusing on a limited number of competencies in order to create value.

The Components of an E-manufacturing Strategy

The potential of e-manufacturing is of a strategic nature and a functional operations strategy will be of concern to most organizations. The possible components of such a strategy can be viewed as a pattern of medium- and long-term decisions. Recent empirical research (Lowson, 2002b) has identified the components of an e-manufacturing strategy as being selected and blended to represent both the resource-based and market-driven views of policy making. (We acknowledge the role of strategic management may not always be rational and planned. Some strategies will be logical and of a breakthrough nature; others will be adaptive, emergent and incremental.) In the case of the former, companies are seen as a collection of resources, rather than holding purely market positions. The market-driven view of strategy formulation, meanwhile, argues that it is not just the industry that is important, but where the organization wants to compete and the nature of the competition.

Previous empirical work (see Lowson, 2002c) has indicated that operations strategies (including those appertaining to production or e-manufacturing) contain various components. These can be conceptualized as core competencies, capabilities, and processes; strategic network relationships; resources; technologies; and key tactical activities to support a particular strategy or positioning. We will return to these components as part of the case study to examine their application in an e-business setting.

E-manufacturing—The Opportunities

An e-manufacturing enterprise will have a number of revenue-enhancing opportunities that will need to be considered in any e-manufacturing strategy.

Direct Sales to Customers. This allows more members of the supply network to have direct contact with the customer in channels not previously open or controlled by intermediaries (a retailer or wholesaler, for example).

Twenty-Four-Hour Access from Any Location. Accessibility, in theory, could be 24 hr a day and 7 days a week for the placement of orders. However, unlike a retailer, the goods are not always provided straight away. Similarly, geographic location is no barrier to accessing e-services, but e-fulfillment may not always be feasible.

Aggregating Information. Sales can be increased by offering information regarding a large selection of products and services that do not have to be held in inventory (a corresponding variety provided by a retailer would entail having a large stock-holding).

Personalization and Customization. The Internet offers the potential to use customer's personal information to intelligently guide each buying experience and increase sales. At an F2B level it is possible to establish customer-specific sites to personalize the buying experience.

Speeding Up Time to Market. Increased revenues are available by introducing new products much faster than using physical channels. A new product can be made available as soon as the first unit is produced.

Flexible Pricing. An e-business can alter prices by changing one entry in a database. Prices can be adjusted rapidly to reflect inventory and demand preferences.

Price and Service Discrimination. Prices can also be altered to reflect the buying power of individuals. Prices can be adjusted for each microsegment and even by individual customer—rather than having a single price for all.

Efficient Funds Transfer. The payment collection cycle can be much shorter for an e-business—this can substantially improve cash flows and possibly increase revenue.

Shipping Time and E-fulfillment. This is still a potential disadvantage for an e-business. For example, physically buying a pair of jeans allows the customer to touch and try on the goods. After payment the customer leaves with her purchase. For an e-business, shipment takes time (unless we are dealing with services and goods that can be downloaded from to a computer). Shipping costs are also a common cause of concern with many consumers abandoning the online "shopping cart" when they add the cost of shipping.

Centralized Inventory. The ability to stock goods in one location rather than spread over numerous scattered locations allows the firm to keep substantially lower amounts by avoiding duplication.

Reduced Facilities Cost. This benefit is obtained by avoiding the need for stores and other outlets in expensive prime locations.

Self-sourcing. Consumers often can do work of the business. For example, delivery services encourage the customer to track their own deliveries, purchase tickets online, and pay bills and set up banking facilities online. These advantages may be the single most important feature of the Internet.

Job Specialization and Scheduling. This allows companies to centralize operations, giving rise to opportunities of greater task specialization. For service firms especially, this is a valuable benefit as it allows a high degree of customization to be retained in the front-end of the service, but this is supported behind the scenes by standardized tasks and processes that can utilize economies of scale through high-volume transactions. The Internet alleviates the fundamental tradeoff between customization and efficiency, especially when applying technology to 24-hr, 7-days-a-week services.

Other Positive Cost Impacts. Cost reduction opportunities include reduced product handling and shorter supply networks, postponement of product differentiation until the order is placed, decreasing delivery cost and time (downloaded goods), reduced physical processing costs, reduced inventory costs, and improved supply chain coordination arising from better information flows.

E-manufacturing—Pitfalls and Challenges

Just as certain opportunities will exist for an e-manufacturing system, there will also be limitations. The operations strategy seeks to minimize and mitigate the following:

Technology. An e-business and its e-manufacturing strategy will live or die by its technological choices. The flexibility brought by information technology will also have a price in terms of the necessary development and maintenance. Often, firms will not possess the expertise necessary in these areas and will either have to outsource or hire the human resources necessary. This can result in higher costs and loss of control. In addition, according to Moore's law, computing power doubles every 18 to 24 months. To keep pace, investment in new technology is both necessary and substantial.

Trust and Security. These remain major concerns for many businesses contemplating an e-manufacturing strategy. The former is fundamental to the proper functioning of e-manufacturing interfaces between firms. Information systems in particular must remain open to enquiry and use by a number of parties to an intranet or extranet. The term "network security" is regarded by many as a contradiction in terms. The use of e-manufacturing increases an organization's visibility on the Internet and more exposure and availability of the product increases turnover and profit. However, it also opens the operation to a wide number of security threats that must be identified and planned for using the necessary security devices.

Increased Shipping. Rather than delivering one shipment to a store for customers to purchase, e-manufacturing must ensure that each individual purchase is delivered. For tangible goods, this has high-cost implications that cannot be avoided—especially if the products are fragile or large.

Accountability/Legality. Accountability for the sale of controlled items is always a problem with the anonymity of the Internet. It is difficult to ensure that items such as liquor or prescription drugs do not fall into the wrong hands. Similar problems arise regarding copyright of music and other intellectual property.

Communication Barriers. Despite the customized Web pages of Amazon and Dell, for example, there is still a barrier to communication without two-way or face-to-face interaction.

Other Negative Cost Impacts. Higher costs are likely in the following: inbound and outbound transportation costs as the e-business will tend to aggregate inventory at a central location rather than at a number of geographic points; increased handling costs if customer participation is reduced (groceries, for example, where some tasks are performed by the customer in store); and, a large initial investment in information infrastructure.

E-manufacturing and the Small- to Medium-Sized Enterprise (SME)

In many developed nations, SMEs make up a large proportion of the manufacturing base. As such, the transformation to a form of e-manufacturing may well be inevitable as customer demand increases for faster delivery and instant access to order schedules. Additionally, when there is a large power imbalance between SME supplier and a larger customer organization these pressures will be exacerbated.

As we shall see later, cost savings are just one of the benefits of e-manufacturing and other electronic processes in "back-end" operations. SMEs need to embrace the latest technologies in order to penetrate nontraditional markets and remain flexible, responsive, and competitive, challenges made more onerous by globalization, liberalization, and technological advancement.

For many firms, including SMEs, the plant floor is often the weakest link in the supply system and this is likely to be the largest gain of e-manufacturing. However, the costs of the necessary technologies may well be prohibitive. Unfortunately, some of the cheaper systems available will be those that fail to integrate with other organizational processes and the supply chain, dissipating the benefits of e-manufacturing.

GLOBALIZATION OF AN E-MANUFACTURING STRATEGY

Although e-manufacturing can be localized, distributed e-manufacturing brings with it certain operations strategic considerations, one of which is location. Theoretically, under such a system, disparate production sites can be controlled from a central location. Yet, there remain a number of other necessary strategic considerations.

Now, even small- and medium-sized manufacturers can source and trade on a global basis. Advanced forms of information communication technology and the Internet have made feasible operations and the selling and buying of products on a global basis. New markets are available in corners of the world previously unthinkable; even micro-organizations can realize benefits from utilizing (often through outsourcing) international resources for both labor and materials.

International competition is a complex arena for any firm. As Meredith and Shafer (2002) suggest, organizations now operate in domestic, exporter, and international markets at the same time. Global firms, joint ventures, partial ownership, strategic alliances, outsourced production and services, build and assemble offshore agreements, foreign suppliers of parts, reimporting manufacturing arrangements, and now e-manufacturing are all commonplace. However, despite global production being used to reduce direct costs, it exacts a heavy price in terms of coordination and flexibility.

Meredith and Shafer suggest that global coordination consists of many elements. An accurate information provision is required in the supply chain to minimize the cost of errors as well as those associated with storage and inventory holding. Indeed, as global coordination becomes more of a problematic issue, many firms will seek to establish closer "sole-source" arrangements with a smaller, "preferred" supply base.

Growth in world trade and the extension of operational activities beyond the boundaries of the firm has clear implications for competitiveness and the dimensions of a manufacturing strategy. Operations have a crucial role to play in international competitiveness. Strategic operations decisions regarding such issues as the responsiveness and flexibility of domestic supply compared with outsourcing to less flexible, low-cost, low-wage producers offshore; the appropriate locations for international facilities; the output capacity of plants; and the labor–technology tradeoffs in each location are necessary. There are six primary characteristics of the transformation system that are crucial when making such decisions (these have been adapted from the work of Meredith and Shafer, 2002):

Efficiency or "doing the thing right"—Measured as output per unit of input. The problem comes in making meaningful comparisons and choosing the right measures for these inputs and outputs.

Effectiveness or "doing the right thing"—Are the right sets of outputs being produced? Are we focused on the right task?

Capacity—The maximum rate of output attainable. Here, a balance must be struck between capacity and efficiency.

Quality—Are the quality levels of the output right and can they be consistently attained?

Response time—How quickly can the output be produced? Or preferably, can the output levels be set to respond to the exact nature of "real-time" demand?

Flexibility—Can the transformation system produce different outputs? How easily? How fast? What levels of customization are possible?

The Right Choice of a Manufacturing Strategy for E-manufacturing

The choice of production strategy will reflect two diametrically opposed variables: the levels of cost reduction necessary, and the levels of flexibility and responsiveness required (Lowson, 2001). Figure 4 demonstrates the manufacturing strategies that might be applied—depending upon the objectives of the enterprise. The matrix in Figure 4 details the two variables. The vertical axis shows the operational cost of a particular manufacturing strategy. The horizontal axis displays the degree of flexibility and responsiveness. The implications for each international strategy are critical for both trade and competitive edge. The various types of strategy can be examined together with their advantages and disadvantages in terms of flexibility, responsiveness, and cost.

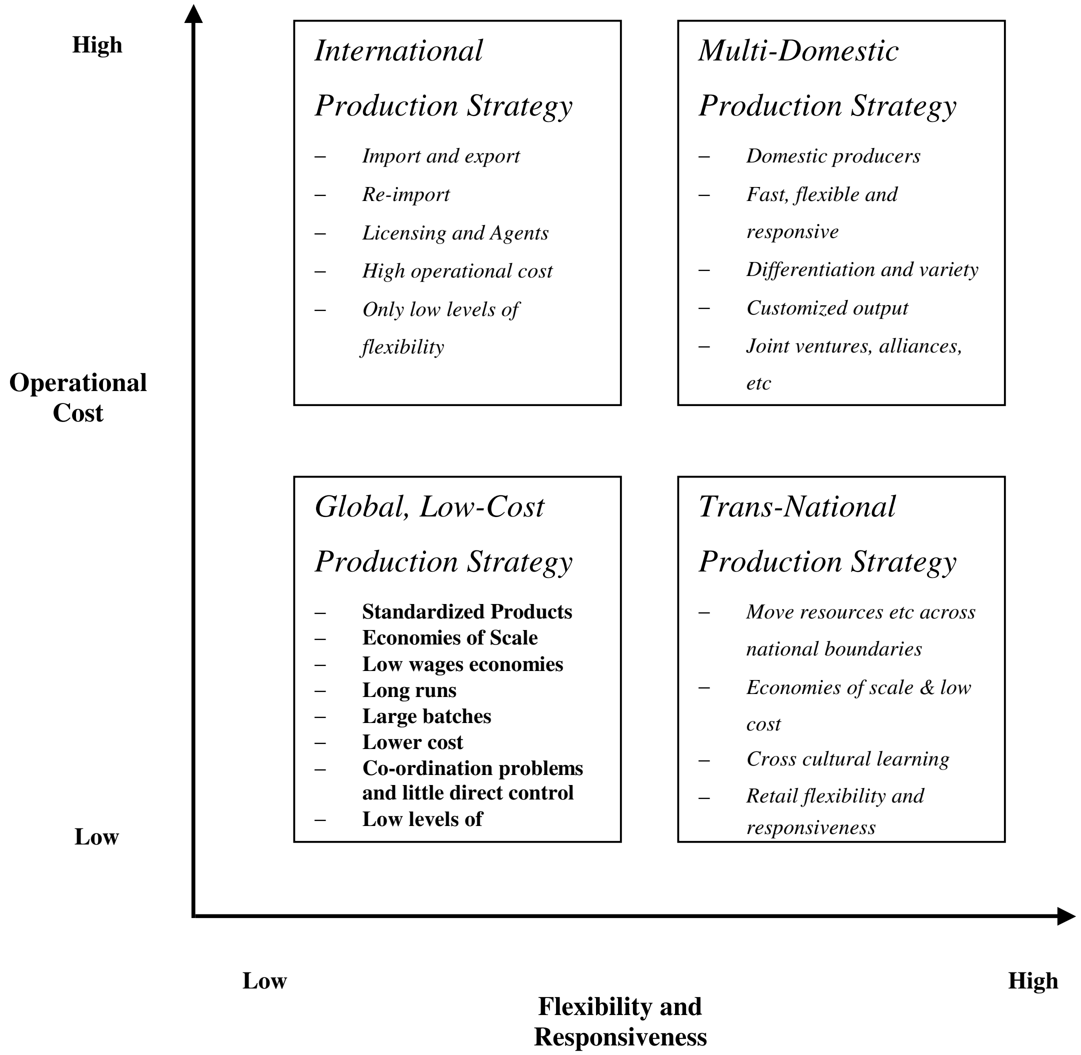

High
Operational Cost
Low
International Production Strategy
– Import and export
– Re-import
– Licensing and Agents
– High operational cost
– Only low levels of flexibility
Multi-Domestic Production Strategy
– Domestic producers
– Fast, flexible and responsive
– Differentiation and variety
– Customized output
– Joint ventures, alliances, etc
Global, Low-Cost Production Strategy
– Standardized Products
– Economies of Scale
– Low wages economies
– Long runs
– Large batches
– Lower cost
– Co-ordination problems and little direct control
– Low levels of
Trans-National Production Strategy
– Move resources etc across national boundaries
– Economies of scale & low cost
– Cross cultural learning
– Retail flexibility and responsiveness
Low
High
Flexibility and Responsiveness

Figure 4: Cost and flexibility considerations in choosing a global production strategy. From Hill & Jones, 1998, and from Hitt et al., 1999.

International Production Strategy

Here, the domestic enterprise imports and exports goods and components from its home country to another—a strategy relying upon the granting of licenses and agents abroad. In this instance, the firm operates in high cost, but also low-responsiveness mode. It is the least advantageous, with low local responsiveness or flexibility and little cost advantage as the transformation processes are some distance from the market. However, as an international manufacturing strategy, it is sometimes the easiest to establish with little change in existing operations and risk exposure.

Multidomestic Production Strategy

In this instance manufacturing decisions are decentralized and taken in each particular country concerned. Organizationally the firms involved are usually strategic business units, subsidiaries, franchises, outsourcing partners, or joint ventures with suppliers. Local producers in local markets maximize response and flexibility and encourage differentiation and variety—often high degrees of mass customization to accommodate local tastes are also an option. Control and coordination are also easier using this strategy as the firm is not just exporting a product but also managing the processes. There are, however, cost disadvantages using this multilocal approach, although flexibility and responsiveness increase with closer ties to the market. The advantages and disadvantages of this type of production strategy can also be applied to sourcing from domestic vendors rather than offshore.

Global, Low-Cost Production Strategy

This strategy is applied to standardized products and processes that can be produced in bulk (often in advance of a sales season) by large manufacturers operating in low-wage economies such as Asia. Economies of scale and low cost are possible due to long runs and size of operations. Coordination is often a problem as control over the numerous offshore suppliers can be difficult; there is also a degree of instability in these arrangements. The strategy also lacks responsiveness and flexibility to changes in demand.

Transnational Production Strategy

This strategy seeks to achieve the best of both worlds. It exploits economies of scale and learning, as well as pressure for responsiveness, by recognizing that core competencies, capabilities, or processes do not just reside in the domestic or "home" country. The strategy is transnational as it moves people, processes, material, and ideas that transgress national boundaries. These firms then pursue differentiation, low cost, and response simultaneously. We can think of such firms as "world" companies whose country of origin is unimportant as they possess an independent network of world-wide operations. Key activities are neither centralized nor decentralized; instead the resources and activities are dispersed, but specialized, so as to be both efficient and flexible (Bartlett & Ghoshal, 1992). In fact these firms are, as Toffler (1994) suggests, "stateless."

In summary, global production strategies increase the challenges and opportunities for most organizations. Many domestic firms have chosen to develop internationally for strategic reasons such as cost and supply network improvement. However, there are also problems in terms of providing sufficiently flexible and responsive operations that can provide the variety of goods increasingly in demand.

FORMULATING AN E-MANUFACTURING STRATEGY

A production or manufacturing strategy will direct the translation of competitive and market aspirations into tangible goods. In other words, having ascertained the value sought and the various order qualifiers and winners, it is necessary to arrive at an operations strategy that will make decisions about various product groups that will be made and delivered.

Developing an E-manufacturing Strategy

The manufacturing strategy cannot be designed in a vacuum. It requires the input of customer needs and requirements as well as resource and capability considerations. The subsequent strategic vision will have to be aligned to the wider business strategy in making decisions about the target market, product groups, and core operational competencies. We must also remember that these decisions are strategic and they will involve radical "step" changes. To create a distinctive or core competence Skinner (1996) suggests that "tinkering with the current system" and just adding new technologies is not enough. In a complex and dynamic commercial environment, new techniques must be replaced by "a whole new product realization system that is different and better than any offered by a competitor."

A strategy for the manufacture of goods is directly concerned with the transformation of raw materials and/or components into products. Often these physical transformations will take place as part of a function (operations or manufacturing) and the operations strategy in this case is at a functional level. The type of strategic decisions to be made can be summarized under the following categories (Harrison, 1993):

How are the various product transformation processes defined?

How are these transformational systems linked?

What are the basic operational principles (processes and infrastructure) for each transformation system?

How is the actual transformation carried out?

What are the quantitative limits of the transformational process?

Where is the process located?

How are the processes physically organized?

Who owns the transformation system?

How are the design attributes arrived at?

How is quality controlled and improved?

How are the informational needs of operations met?

How are human resource needs met?

Hill (2000) provides an important and explicit framework for a manufacturing strategy that reflects the close association between manufacturing and corporate decisions. His approach in developing a strategy places a strong emphasis upon strategic integration, in particular, the links between manufacturing and marketing essential in an e-business environment. In Figure 5 (adapted from Harrison, 1993) we can see how these various steps form a strategic framework that provides a close association between strategic thinking and various choices necessary. The various steps are now described in more detail:

Step 1—Corporate Objectives

The objectives for the business are the basis for establishing clear, strategic, directions with an awareness and willingness to succeed. The directions also define the boundaries and parameters against which the various inputs can be measured and established—a coherent corporate plan. Of course, these objectives will vary from one enterprise to another and will mirror the economy, markets, opportunities, and preferences involved in the situation. These broad objectives may include growth aims, survival, profit expected, return on investment, and other financial measures. However, they can also reflect other concerns such as environmental resource issues and employee policies.

Step 2—Strategic Positioning

Any commercial enterprise will need to adopt a particular strategic positioning that will underpin both its corporate objectives (Step 1) and marketing strategy (Step 3). We can immediately see the difficulty of a linear approach in this. Positioning can be conceptual (variety-based, need-based, and access-based—or a combination of all three); or practical (based upon cost, quality, flexibility, response, etc). This positioning involves a set of decisions that will take into account strategic priorities but will also dictate the operational activities that will be necessary to achieve these priorities.

Step 3—Marketing Strategy

The marketing strategy is just one of the many functional strategies the organization may choose to adopt. These

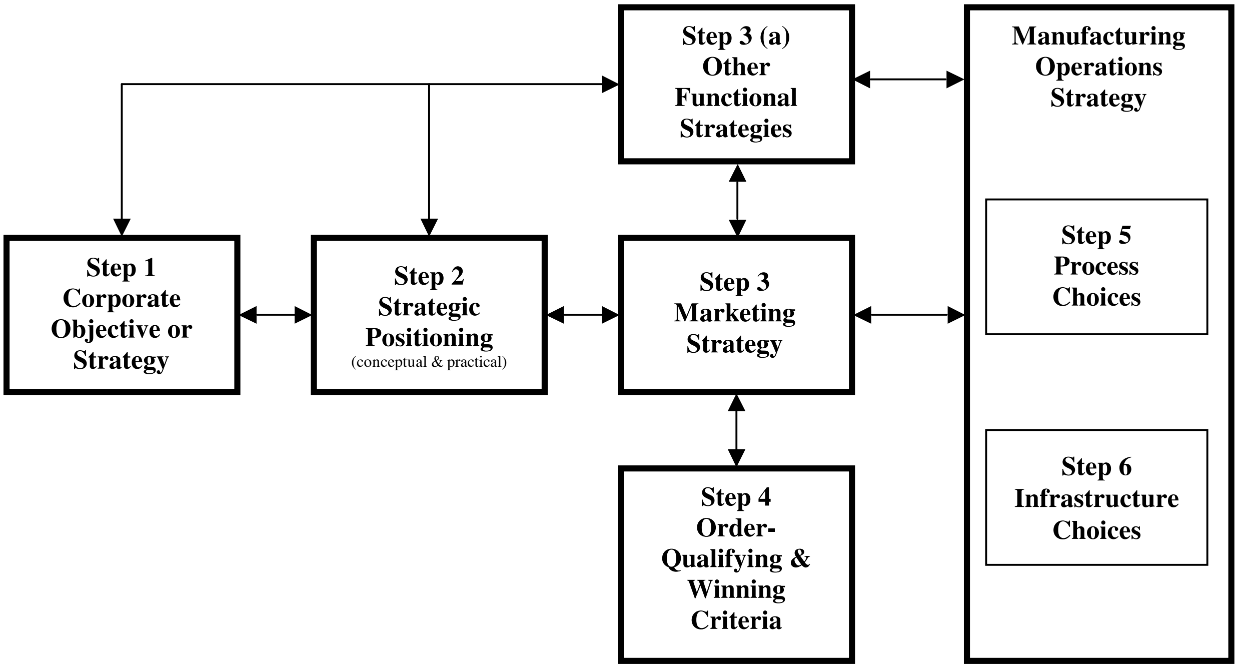


Figure 5: Manufacturing strategy framework.

include, for example, the operations or production strategy, financial strategy, human resource management strategy, and information systems and information technology strategy.

The marketing strategy will include the following:

Marketing planning and control, i.e., establishing the target markets for various product groups;

Analysis of product markets to estimate value sought, trends, buying behavior, and competitive activity; and

Identification of target markets and agreeing on objectives for each, both short and long term, and the action plans necessary.

In addition, the marketing plan will need to establish levels of support and the various resources necessary to meet objectives. The output of the marketing strategy will determine and identify product markets and segments; the range, mix, and volumes of products involved; the various product attributes necessary to satisfy value demands; and the overall marketing strategy for the various target markets (for example, a leader, aggressive competitor, new player, or follower).

Step 3(a)—Other Functional Strategies

In parallel to the marketing strategy, other functional strategies will need to be developed (too often, strategic planning fails to take this into account sufficiently early in the process). These functional strategies will cover a number of areas, but for our purposes, the manufacturing strategy is most important. Here, it will be important that the strategy both supports the corporate and marketing strategies and be compatible with other functional strategies. The various decisions about, and strategic management of, core competencies and processes, technologies, resources, and key tactical activities used in the organization and its supply network must reflect the value demanded by a customer/consumer, as well as the strategic priorities of marketing and the business as a whole. The resultant production strategy blends the various components into a fusion or strategic architecture with various levels of emphasis. This approach allows each strategy to be customized and to fit the requirements of not only the marketing strategy, but also the target markets for each product or customer group (Lowson, 2002a).

Step 4—How Products Qualify and Win Orders in the Market

The task of the manufacturing strategy is twofold. First, it is to meet the criteria necessary to qualify for the market through operational effectiveness (*performing similar activities better than rivals*). Second, it is to win orders and sustainable competitive advantage through differentiation from competitors (*performing activities different from rivals' or performing similar activities in different ways*). The qualifying criteria are the essential features of a product while order- winning criteria provide the basis for customer choice between alternative offerings. The following are some of the possible customer choice criteria that, in differing circumstances, or product life-cycle stages, may be associated with either market qualification or order-winning: total price, specified product or service attributes, quality of the product, design and design flexibility, delivery arrangements, delivery performance, volume and mix flexibility, speed of quotation service, product features, after-sales support, and the stability of continuing customer support.

The decisions in this step are strategic and need to be taken at a cross-functional level. They are not the prerogative of one function, as there will be many conflicting perspectives of a firm's products and markets. This underlines the strong link necessary between corporate marketing proposals and commitments and the manufacturing processes and infrastructure necessary to support them.

Step 5—Manufacturing Strategy (Process Choice)

There will be number of alternative processes available to make particular products. In making such choices, the key considerations will include volume, degree of customization, mix, complexity in the production process, and the necessary order-winning criteria. However, each type of choice involves both current and future tradeoffs, often between cost and variety.

Process choice relates to the basic mode of operation of a system, e.g., producing one-off projects designed for a specific customer, at one extreme, and continuous production, of say, petroleum or chemicals, at the other. The process choice will involve decisions regarding the use of alternative processes, the tradeoffs embodied in the process choice, the role of inventory in the process configuration, decisions on whether to "make" internally or "buy" parts or components externally, and capacity decisions (size, timing, location).

Step 6—Manufacturing Strategy (Infrastructure Choice)

This concerns the non-process features within the production strategy, and can include the procedures, systems controls, compensation systems, work-structuring systems, and organizational issues within manufacturing. Infrastructure choices include the necessary support needed for the operations function, planning, and control systems; quality assurance, management, and control systems; manufacturing systems engineering, data, and information requirements; clerical and informational procedures; compensation agreements; and work and organizational structures.

COMPETITIVE ADVANTAGE FROM AN E-MANUFACTURING STRATEGY

A summary of the advantages of e-manufacturing and its potential for a strategic competitiveness can now be offered. There are a number of unique opportunities offered by e-manufacturing:

Customized design—Customers can electronically transmit requirements to remote locations and become closely involved in the design of individualized goods. The final product can be quickly manufactured and distributed.

Greater product customization—As described above, it is a major advantage in terms of manufacturing to order unique goods, at relatively low cost.

Alliances—Organizations can use electronic linkages (see Figure 1) to form alliances and rapidly bring on-stream new products using advanced manufacturing techniques such as quick response and agile and virtual manufacturing.

New and modified products—The formation of strategic alliances takes advantage of brief windows of opportunity in the market. The blending of core competencies in such alliances, even on a virtual plane, has substantial implications for reducing time in product life cycles.

Increased market share—SMEs, for example, can more easily advertise their capabilities using computer networks and tender for work on projects required by other firms as part of an outsourcing strategy.

Technological applications—Software system brokers can connect users to sophisticated manufacturing tools for a temporary access to technology normally too expensive to acquire.

Procurement—Intelligent procurement systems can facilitate and speed parts procurement, billing, and payment transactions. This will reduce cost, improve accuracy, and meet customer needs in a timely manner. The ability to link efficient procurement strategies and the business workflow together with robust technological implementation will differentiate market leaders.

Production processes—E-manufacturing also accommodates swift changes in product designs, product lines, and the processes necessary.

Cost reduction—It now seems widely accepted that e-enabled production facilities have the strategic potential to reduce costs and improve efficiency. A new generation of Internet-compatible equipment will make it possible to use monitoring systems to identify cost-effective locations for large orders and ensure seamless flow of information from the factory floor.

Product information—Packaging information, providing access to manufacture and product information, and informational e-tools are becoming critical differentiators in the customer's purchase decision.

Consumer contact—Perhaps one of major advantages of e-manufacturing will be its ability to bring the consumer closer to the producer. In recent years, retailers have held power due to their closeness to the customer and ability to capture and manipulate customer data (mainly using point-of-sale scanning systems). This not only gives information regarding preferences, it also allows more accurate replenishment and re-estimation and reorder systems. The advent of e-manufacturing establishes a direct information link between the consumer and manufacturer without any retail intervention.

This chapter concludes with an empirical case study, which gives some indication as to the application of the Internet and the potential of e-manufacturing.

CASE STUDY APPLICATION

Pentwyn Splicers—E-manufacturing for the Small- to Medium-Sized Enterprise

> Just a year ago, the small engineering firm Pentwyn Splicers was on the edge of bankruptcy. Now it is turning itself around thanks to innovation, the Internet and e-manufacturing.

> Hard work, new product ranges and the development of an acclaimed Internet site have lifted export sales by more than a fifth since its black days. Turnover, helped by fresh markets in Thailand, Korea, Poland and Indonesia, has reached $600,000 at the company, which makes devices known as splicers for the textiles industry. Last year the website won the final of the ISI/InterForum E-Commerce award. Johnson (1999)

The pneumatic splicers made by this company allow weavers to join two separate lengths of yarn without tying knots, which introduce ugly lumps into the fabric and potentially jam up looms. These simple yet innovative products have given Pentwyn, an SME employing 12 staff, a way of surviving the huge slump in the textile sector. However, the firm also had to develop new markets, and the Internet and e-manufacturing proved the solution.

Pentwyn wanted to improve its customers' experience by providing better quality, faster information on its products, and a better response to orders. It needed to reach new markets in a cost-effective manner. The company also had to cut costs, particularly the cost of traveling to service customers and the costs of developing new products.

The firm incurred some cost in hardware and software and the investment time and resources in setting up a Web site and its e-manufacturing function was considerable. It was estimated that approximately 300 hours over a 12-month period were devoted to this (many by specialist engineers and designers). The company introduced e-mail to aid communication with agents and customers and established a Web site to market its products to new customers. Having Internet access also allowed it to source its business needs online.

Pentwyn increased export turnover by 20% in its first year of introducing these new technologies into the business. Its Web presence has enabled it to sell to new customers and markets it could never previously reach (Poland and Thailand, for example). It has improved the speed of communications with customers and is now able to send high-resolution video stills of products and full-technical manuals by e-mail, at much reduced cost.

By using the Internet as a market communications device, prospective customers can seek information and send an e-mail to which the company responds product photographs attachments. For example, a firm in Estonia recently sent samples of their yarn and the company responded immediately by e-mail, attaching photos of splicers joining the ends together. The company also delivers manuals and product updates electronically to clients: "It saves massive printing costs," according to Graham Waters, Managing Director (Waters, personal communication, June 15, 2000). Orders received automatically trigger production planning and the sourcing of materials. Customers can also track the progress of the order in production at a distance and even make some last-minute adjustments to volume or mix. For larger accounts, Pentwyn allows electronic access to inventory levels to provide automatic replenishment orders (as part of their vendor-managed inventory system), which in turn prompts further production scheduling. In addition, customers and suppliers also participate in Internet-based product design support systems.

More than 25% of Pentwyn's new business can be tracked directly to the introduction of the Internet. It made a 10% saving in cycle-time necessary for product development by sourcing materials from new suppliers discovered via the Internet. Online searches have also cut the cost of foreign travel. Overall, the firm estimates that the move toward e-business and e-manufacturing has enabled a cost saving in the region of 10–15%. Further research conducted with the firm demonstrated that the introduction of e-manufacturing had substantially reduced product development time and the levels of inventory and work-in-progress. It was estimated by the managing director that these moves would result in a further 15% saving in both direct and indirect cost.

In the future, the company plans to set up a password-protected extranet for agents and existing customers to improve the level of service that can be offered. It also hopes to translate its online manuals and price lists into French, German, Spanish, and Italian. According to Waters,

> We have a simple vision of the future. We want our web presence to be like a bookshop. We want visitors to stay for ten minutes and browse, not be pressurized and not threatened by hard sell. The real revelation to us is the progress achieved through e-mail communication and website development. The quality of our responses, and their speed, has transformed relationships with customers and our agents. E-manufacturing also offers great opportunities to all small manufacturing companies to transform their approach to business, and the quality of their service. We have found it an ideal medium to equip us to expand in a global market by allowing customers a direct input into producing customized goods and services. But, firms cannot just rely on the technology alone. Their operational systems, and as we have now learned, operations strategies, have to also be developed to support e-commerce. (Waters, personal communication, June 15, 2000)

It is possible to summarize the main components of Pentwyn's e-manufacturing strategy using the building blocks introduced in the previous sections.

Core Competencies, Capabilities, and Processes

These can be further subdivided as follows:

Process-Based (Derived from Transformation Activities). For Pentwyn, the main transformation involved in their e-manufacturing systems concerned information. The core competence they developed involved producing information regarding products and services accessible on an international basis, seeking information from potential customers and suppliers globally, and then matching the two streams by providing business opportunities. In addition, the automatic conversion of this information into signals that triggered production, call-offs from inventory, supplier scheduling, and replenishment activities could all be shared with major customers and suppliers.

System- or Coordination-Based (across the Entire Operation System). The information transformation described above had to be coordinated across their current operations to manufacture for new customers or source from new suppliers. The ability to link and manage the various data streams to control manufacturing schedules, raw material supply, inventory levels, and order processing is of key importance in an e-manufacturing environment.

Organization-Based (across the Entire Organization). For an e-business, organizational structures, processes, and procedures will need to radically change. First, competencies and processes will have to reflect a new tempo of immediate operation. With a faster, more flexible, "real–time" response, Pentwyn had to develop the ability to process information and react to demand across the whole organization much faster than before. In addition, their domain of operation became global and much larger. This involved developing new capabilities and competencies in what was now an outward-facing firm.

Network-Based (Covering the Whole Supply Network). As indicated above, the ability to manage a wider network of operations (both suppliers and customers) quickly became a core necessity for Pentwyn, as did the integration tasks involved in developing a supply network. In an e-business and e-manufacturing domain, greater reliance is placed upon external relationships.

Resources (Individual Resources and Their Unique Combination)

Tangible. Clearly the technologies used by Pentwyn in their e-manufacturing strategy proved vital. These included ISDN lines, CAD software, integrated numerically controlled machinery, sufficiently powerful stand-alone PCs, Internet access, e-mail, video-conferencing facilities, and membership of an electronic trading system. Other enabling techniques and implementation issues included Java programming, 3D modeling and simulations, open computing, and interactive design.

Intangible. The various communication and information systems developed for coordination became an important contributor to Pentwyn's success. They had to develop new business systems for a new method of trading. Thus, the skills in developing new management structures became fundamental.

Human. Specialized skills had to be quickly developed to take advantage of the e-manufacturing opportunities. Knowledge, motivation, and training all played an important part in providing a unique resource combination, capable of utilizing the e-business environment.

Technologies

Over and above resource technologies used, as with any e-business, future technological process development and application will be vital. The innovation and know-how to apply technological advances to new products and processes becomes as important as the technology itself.

The Key Tactical Activities

These are the necessary tactical activities vital to Pentwyn in supporting their business strategy and strategic positioning. Although not strategic in themselves, e-manufacturing decisions had to be made regarding the future of these aspects, e.g., fast and effective sourcing, shared product design, higher quality levels despite speed of response, information system integration, close communication and working partnerships with suppliers and customers, and supply system integration.

The various building blocks of any manufacturing strategy will need to be fused into a particular strategic architecture that reflects the importance placed upon some elements vis-à-vis others. In so doing, individual strategies are developed with a unique emphasis that reflects the commercial situation (Lowson, 2002a).

CONCLUSION AND FURTHER RESEARCH

It is clear that e-manufacturing has huge potential with benefits extending beyond the ability to monitor production systems at a distance. The strategic implications are perhaps most noteworthy in providing a clear route to increased flexibility and responsiveness as well as the ultimate goal of mass customization and product individualization. However, the full benefits of e-manufacturing require high degrees of coordination; both internally and externally. At present, many organizations struggle with these implications. Often, production and supply chain processes and information communication systems lack the integration necessary to take full advantage of the e-manufacturing potential.

It is suggested this lack of integration poses one of the most fruitful areas for future research. In addition to the technical barriers, many of which can be surmounted, the main challenge remains cultural. The true potential of e-manufacturing cannot be achieved without an open culture of shared communication between organizations in a supply network. This enables the coordination not only of production processes, but also of other equally important aspects of order fulfillment. Unfortunately, many firms, and indeed whole industries, still struggle with this concept despite so-called "preferred" supplier agreements, alliances, and partnerships. Competition may now be at supply chain level, but the very fact that a competitive ethos still persists between enterprises at different nodes of a supply system poses an immense barrier to full e-manufacturing deployment.

GLOSSARY

E-manufacturing Electronically mediated data/information exchanges, and the information communication technology systems that support them, to facilitate the operation, enhancement, reengineering, and integration of the range of *production* processes, both within an organization and externally (up- and downstream).

E-business All electronically mediated information exchanges, both within an organization and with external stakeholders, supporting the range of business processes.

Flexible response The ability to satisfy demand in all its many forms.

Manufacturing strategy Patterns of decisions that shape the long-term capabilities of the organization to provide products and the contribution to the overall corporate or business strategy.

Supply network An interconnection of organizations using different processes and activities to add value.

CROSS REFERENCES

See *Electronic Commerce and Electronic Business; Feasibility of Global E-business Projects; International Supply Chain Management; Inventory Management; Supply Chain Management; Virtual Enterprises.*

REFERENCES

Barlett, C., & Ghoshal, S. (1992). *Transnational management*. Chicago: Irwin.

Harrison, M. (1993). *Operations management strategy*. London: FT Pitman.

Hill, T. (2000). *Operations management: Strategic context and managerial analysis*. Basingstoke: Macmillan Business.

Hill, C. W. L., & Jones, G. (1998). *Strategic management*. New York: Houghton-Mifflin.

Hitt, M., Ireland, R. D., & Hoskisson, D., (1999). *Strategic management*. Cincinnati, OH: Southwestern College Publishing.

Johnson, A. (1999, February 3). Firm on the abyss spins recovery from the Web. *The Express*. 7.

Lowson, R. H. (2001). Assessing the true operational cost of offshore sourcing strategies. *International Journal of Logistics Management, 4*(3), 27–51.

Lowson, R. H. (2002a). *Strategic operations management: The new competitive advantage*. London: Routledge.

Lowson, R. H. (2002b). Operations strategy: Genealogy, classification and anatomy. *International Journal of Operations & Production Management, 22*(10), 1112–1129.

Lowson, R. H. (2002c). Strategic operations management—The new competitive advantage. *Journal of General Management, 5*(2), 89–112.

Meredith, J. R., & Shafer, S. M. (2002). *Operations management for MBAs*. New York: Wiley.

Pires, J. N. (2001). EmailWare: A tool for e-manufacturing. *Assembly Automation, 21*(2), 129–235.

Porter, M. E. (2001, March). Strategy and the Internet. *Harvard Business Review,* 63–78.

Skinner, W. (1996). Three yards and a cloud of dust: Industrial management at the century end. *Production and Operations Management, 5*(1), 11–12.

Toffler, A. (1994, March). Recipe for Intelligence. *Information Today,* 61–63.

Toussaint, J., & Cheng, K. (2002). Design ability and manufacturing responsiveness on the Web. *Integrated Manufacturing Systems, 13*(5), 328–339.

Van Hoek, R. (2001). E-supply chains, virtually non-existing. *Supply Chain Management: An International Journal, 6*(1), 21–28.

FURTHER READING

Advanced Manufacturing (2003). Retrieved January 10, 2003, from http://www.advancedmanufacturing.com

AMR Research (2003). Retrieved January 3, 2003, from http://www.Amrresearch.com

Bury, S. *Internet technologies: Interconnectivity the key to wired future*. Retrieved November 9, 2002, from http://www.advancedmanufacturing.com/technologies.htm

E-manufacturing (2003). Retrieved November 9, 2002, from http://www.advancedmanufacturing.com/industrytest.htm

Industry Week (2003). Retrieved January 6, 2003, from http://www.industryweek.com

Mathieu, R. G. *Manufacturing online*. Retrieved November 9, 2002, from http://www.advancedmanufacturing.com/manufacturing.htm

Trombly, R. (2000). E-business models. *Computerworld, 34*(49), 61.

Extensible Markup Language (XML)

Rich Dorfman, *WebFeats! and Waukesha County Technical College*

WHAT IS XML?

XML (extensible markup language) is a text-based, human- and machine-readable language used for sharing data over the Web, intranets, and extranets. XML is platform- and language-independent, making XML to data what Java is to platform-neutrality and .NET is to language-neutrality. XML is the *lingua franca* of data.

XML is three things:

XML is a core language, defined in a World Wide Web Consortium (W3C) Recommendation (World Wide Web Consortium, 2000), that is used to organize data in a hierarchical structure.

XML is a family of technologies. W3C has issued Recommendations and Working Drafts for several languages that are often used in conjunction with the core language to accomplish frequently demanded tasks.

XML is a standard that defines syntactic and semantic rules for creating XML-based languages. XML-based languages are also known as *dialects* or *grammars* of XML.

Figure 1 shows the XML core language, other languages in the XML family of technologies, and some of the XML-based languages that are built upon them.

The XML core language is similar to HTML in syntax but different with regard to functionality. XML is a markup language because it uses tags to give structure to data. XML is extensible because it enables users to create their own collections of tags, as opposed to HTML, which has a fixed set of tags.

The core language has found and continues to find an ever-increasing range of applications, often in conjunction with other XML dialects. Some of the more popular uses of the XML core language include the following:

Displaying data in a browser or other user agent.

Specifying configuration information for stand-alone applications and servers.

Sharing data between different applications, platforms, and companies.

In this chapter we focus primarily on the XML core language, but bear in mind that its syntax and semantics form the basis of all other XML-derived langu ages.

The XML family of technologies arose as W3C continued to build upon its original XML Recommendation. Some of the members of the XML family of technologies include the following XML dialects:

XSL (XML stylesheet language), a native XML replacement for CSS, is used for rendering XML documents in accordance with a style sheet.

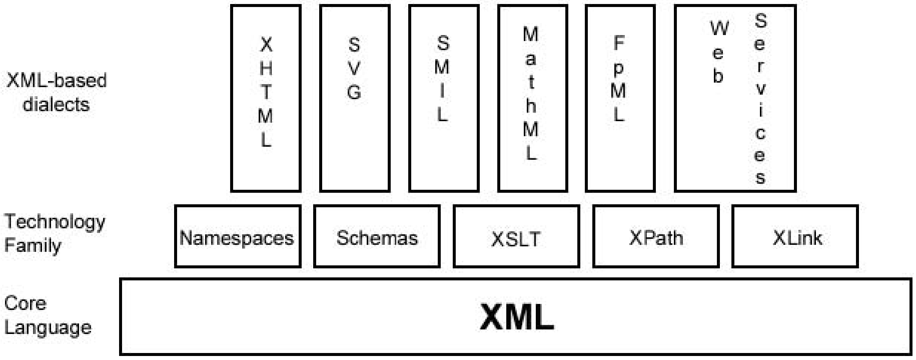


Figure 1: The XML pyramid.

XSLT (XSL transformations), and its companion XPath, are used to transform XML into another format, e.g., XHTML or SVG, for the purpose of presentation.

XLink and XML Base describe a standard way to add hyperlinks to an XML file.

XPointer is a syntax for pointing to parts of an XML document. An XPointer is similar to a URL but instead of pointing to a Web address, it points to pieces of data inside an XML file.

XML Namespaces provide a means to avoid name conflicts when using multiple XML documents within an application.

XML Schemas 1 and 2 help developers precisely define data structures to be used in multiple XML documents. A native XML replacement for DTDs, XML schemas are used to ensure the integrity of shared data.

While W3C has created several XML-based languages, independent companies, industry groups, and academicians have also been active in creating their own specialized grammars. Some examples of XML-based languages include the following dialects:

XHTML (extensible hypertext markup language) features the semantics of HTML 4.01 using the syntax of XML.

MathML (mathematical markup language) provides a grammar for easy rendering of mathematical symbols and equations.

CML (chemical markup language) is an XML vocabulary for representing and drawing molecular information.

MusicML defines a language for rendering sheet music.

FpML (financial products markup language) is a standard metalanguage for describing swaps and derivatives.

HOW DOES XML WORK?

To understand how XML works, we first compare and contrast it to HTML. Next we look at a simple XML example. Finally we discuss XML parsers, the magic that makes it all work.

XML Is Like HTML, Only Different

To understand how XML works, it is helpful to compare it to a language that is familiar to many of us—hypertext markup language (HTML).

HTML consists of tags, or *elements,* that we use to surround blocks of text in an .html file. These tags provide formatting information. When we load an .html file into a Web browser, the browser displays, or *renders,* our text in the format specified by the tags. This bit of magic is made possible by a *parser*—all Web browsers contain their own parsers, or rendering engines, that interpres HTML tags and translates tag formatting information into a graphical display.

HTML tags may also include *attributes* and, if included, each attribute is assigned a *value*. For example, the HTML code fragment `<p align="center">This is a paragraph.</p>` declares that the text "This is a paragraph." is to be formatted as a single paragraph horizontally centered in the browser window. `"<p>"` is the paragraph tag, `"align"` an attribute, and `"center"` its value. The closing `</p>` tag delimits the text enclosed by the tags.

XML uses a similar syntax that includes elements, attributes, and values. But whereas HTML is used for *displaying* information, XML is used to *organize* data in a hierarchical structure. Whereas HTML has a fixed and finite set of tags, XML lets us create whatever tags we wish as long as we conform to its semantic and syntactic rules. Table 1 summarizes similarities and differences between HTML and XML.

A Simple XML Document

Listing 1 shows a simple XML document that structures a company's employee and payroll data. Line numbers appear for reference only and are not part of the actual file.

Listing 1: File payroll.xml.

```
<?xml version="1.0"?>
<!-- payroll.xml -a simple XML file -->

<payroll company="Acme">
  <employee id="23">
     <first_name>John</first_name>
     <last_name>Doe</last_name>
     <salary>$34,000</salary>
  </employee>
  <employee id="24">
     <first_name>Jane</first_name>
     <last_name>Smith</last_name>
```

Table 1 Comparison of HTML vs. XML

FEATURE	HTML	XML
Uses	Used to format text, images and other media for rendering and display in a browser or other user agent	Used to organize data in an hierarchical structure
Has tags (elements), attributes and values	Yes	Yes
Has a fixed, finite set of tags	Yes	No
Has a strictly enforced set of semantic and syntactic rules	No, laxly enforced to varying degree depending on browser	Yes
Can be created in a text editor	Yes, HTML files are plain text	Yes, XML files are plain text
Is contained within a file	Yes, HTML code resides in an.htm or.html file.	Yes, XML code resides in a.xml file but may additionally exist only temporarily as a structure in memory

```
13     <salary>$42,000</salary>
14   </employee>
15 </payroll>
```

Line 1 is the *XML declaration*. Technically, XML documents do not need to start with the XML declaration, but W3C recommends it (Holzner, 2001).

Line 2 is a *comment*. Note that XML uses the same syntax for comments as HTML.

Line 3 is blank, or *white space*. It is included simply to make the code more readable.

Line 4 is the *root element*. The root element is the first element in an XML document and contains all the other elements nested between its start and end tags. The root element may include an *attribute*. Here, the attribute "company" has the value "Acme."

Line 5 begins an element ("employee") that contains an attribute-value pair (id="23"). Note that in the strict XML hierarchy, there are parent–child relationships between all elements. Here, <payroll> is the parent, <employee> the child.

Lines 6–8 contain three more elements, all children of <employee>.

Line 9 closes the <employee> element begun on Line 5.

Lines 10–14 describe another employee of the company. More employees could easily be added by following the same structure.

Line 15 closes the <payroll> root element begun on Line 4 and ends the document.

So now that we have an XML file, what good does it do us? The short answer is, "none." The longer answer is, "plenty, given the right tools to further process the file or format it for display." Our simple XML file is in fact a *data source that is accessible programmatically*. We can query, manipulate, and display XML data using any of several programming languages, in conjunction with an *XML processor*, or parser—the application that reads and interprets our XML.

Parsers

We could of course write code ourselves to parse the contents of an XML file—after all, it is just a text file. However, there is no need to reinvent the wheel, as free parsers are readily available for download. Most likely, we one or more parsers came with software that is installed on our systems.

Parsers have the intelligence to read, write, and manipulate data in XML files, but they have one limitation—they cannot read our minds. To tell a parser what to do, we use an API (application programming interface). Many parsers come complete with their own API tools. Alternately, we might use one company's parser while accessing it via an API from a different party.

We will discuss parsers in more detail later in this chapter. First we need to take a look at *well-formed* and *valid* XML documents—the language's syntax and semantics.

CREATING WELL-FORMED XML DOCUMENTS

What Are Well-Formed XML Documents?

Definition: XML with correct syntax is well-formed XML.

Parsers check XML documents for well-formedness. If a document is not well formed, the parser discontinues processing and returns an error message that indicates where the malformed syntax occurred.

Rules for Well-Formed XML

A well-formed XML document must conform to W3C semantic and syntactic rules (W3C, 2000). The following rules specify the conditions an XML document must conform to to be considered well formed.

Semantics

W3C specifies that a well-formed XML document has three parts: a prolog (which can be empty), a root element, and various miscellaneous components.

The Prolog. The prolog comes at the very beginning of an XML document. The prolog can contain an XML declaration, processing instructions, a document type definition (DTD), and comments. A document does not absolutely require a prolog to be considered well formed, but W3C recommends including at least an XML declaration. If an

XML declaration is included, it must appear on the very first line.

An XML declaration looks like this:

```
<?xml version="1.0"?>
```

To date, the version attribute is always equal to "1.0," although W3C issued a Candidate Recommendation of the XML 1.1 specification in October 2002 and expects to issue a finalized Recommendation in 2003.

Other optional attributes of the XML declaration include the following:

Encoding—specifies the character set used in the document. Defaults to UTF-8. A document may also use Unicode sets such as UCF-2 and UCF-4 or ISO character sets such as ISO-8859–1 (Latin-1/West European).

Standalone—specifies whether the document requires an external file. Setting standalone to "yes" means that the XML document does not require an external DTD or schema or any other external file. We would set this to "no" if you were referencing an external DTD or schema file. DTDs and schemas are discussed in the next section of this chapter.

Besides the XML declaration, the prolog may contain additional *processing instruction*s. Processing instruction tags start with `<?` and end with `?>`. Processing instructions are directives to the *XML processor,* or parser—the application that reads and interprets our XML. Any processing instructions that we use must be understood by our parser; they are not built into the XML Recommendation. Here is an example of a commonly used processing instruction that links a stylesheet to the XML document. It is understood by parsers in both Internet Explorer 5 and Netscape 6:

```
<?xml-stylesheet type="text/css"
  href="myStyle.css"?>
```

The prolog may contain three more things: *comments* and *white space,* which are discussed in Syntax, below, and the *document type definition (DTD),* which is discussed in the next section of this chapter.

The Root Element. The first element that comes after the prolog is known as the root element. All XML documents must have root elements.

The root element consists of a tag pair and includes everything between its starting tag and its ending tag. All other elements in the XML document are nested between the starting and ending tags of the root element. In the simple XML example shown at the beginning of this chapter, `<payroll>` is the starting tag of the root element and `</payroll>` is the ending tag. It's like saying, "This document is a payroll."

Miscellaneous Parts. Optional miscellaneous elements in an XML document may consist of comments, processing instructions, and white space.

We have already met processing instructions in the prolog. There are other processing instructions that may be used anywhere in an XML document. Again, the constraint is that the processing instruction must be understood by the parser.

Syntax

W3C specifies that a well-formed XML document must conform to the following syntax rules.

Tags are delimited by "greater than" and "less than" brackets.

Element tags start with < and end with >.

Elements consist of a starting tag, an ending tag, and everything in between.

Just as the root element must have a closing tag, so too must all other elements. The closing tag consists of the name of the opening tag, prefixed with a slash ("/"). Attributes, if present in the opening tag, are not repeated in the closing tag. For example,

```
<employee id="24">
   <salary>$42,000</salary>
</employee>
```

If the element does not contain text or other elements, we may abbreviate the closing tag by simply adding a slash ("/") before the closing bracket in our element. For example,

```
<payroll>
   <employee id="24" />
   <employee id="807" />
</payroll>
```

Here, `<employee id="24"/>` is a so-called *empty* element, as is `<employee id="807"/>`.

Elements must be properly nested.

In the strict XML hierarchy, there are parent–child relationships between all elements. The root element may contain one or more child elements. Each child element, in turn, may act as parent to its own children. Child elements must be correctly nested within their parent elements. For example, in HTML we can get away with this:

```
<b>
   <i>This text is bold and italic</b>
</i>
```

But XML requires that we close child elements before closing their parents:

```
<b>
   <i>This text is bold and italic</i>
</b>
```

XML tags are case sensitive. Whereas `<employee></employee>` is well-formed, `<EMPLOYEE></employee>` and `<Employee></eMPLOYEE>` are not.

Attribute values must be enclosed in quotation marks. `<employee id="24"/>` is well-formed, but `<employee id=24/>` is not.

Comments in XML. The syntax for writing comments in XML is the same as that used in HTML. For example,

```
<!-- payroll.xml - a simple XML file -->
```

XML preserves white space within elements.

HTML ignores white space. An HTML statement such as

```
<p>Hello.        How are you?</p>
```

looks like this when rendered by a browser:

Hello. How are you?

XML parsers, however, preserve the element's white space:

Hello.   How are you?

On the other hand, parsers ignore the use of vertical and horizontal white space between elements to make code more readable. For example, in

```
<payroll>
   <employee id="24"/>
   <employee id="807"/>
</payroll>
```

XML converts CR/LF to LF.

In Windows applications, a new line of text is stored as two characters: CR LF (carriage return, line feed). In Unix applications, a new line is stored as a single LF character. Some applications use only a CR character to store a new line. In XML, a new line is always stored as LF.

Only five entity references are predefined.

Like HTML, XML uses *entities,* or escape sequences, to allow us to use special characters within XML elements. For example, say we need to use a numerical relationship within an element, such as

```
<relationship value="a > b">
```

We are trying to say, "a is greater than b," but we run into a problem because XML parsers interpret the "greater than" symbol (">") as the end of an element. Instead, we have to use the following entity reference:

```
<relationship value="a &> b">
```

Here, `"&>"` is the entity reference for the "greater than" symbol.

In XML, entity references always start with an ampersand (`"&"`) and end with a semicolon (`";"`). The W3C spec defines, and XML parsers recognize, only the following five entity references:

```
&amp;     The & character.
&lt;      The < character.
&gt;      The > character.
&apos;    The ' character.
&quot;    The " character.
```

If we wish to use any other entity references, we must define them in a DTD or Schema.

Data contained within XML elements are either CDATA or PCDATA.

PCDATA (parsed character data) are just normal, everyday data, such as the "$34,000" in `<salary>$34,000 </salary>`. Data contained within element tags are always considered PCDATA unless specifically declared as CDATA. PCDATA are text that will be parsed by a parser—any tags inside the text will be treated as markup and entities will be expanded.

CDATA (character data) are text that will NOT be parsed by a parser. Tags inside the text will NOT be treated as markup and entities will not be expanded. CDATA are typically used for large blocks of text. Their syntax looks like this:

```
<employee>
<![CDATA[
   This is an unparsed character data area.
     Because it's
   not parsed, I can include < and > symbols
     without having
   to worry about the parser mistaking those
     symbols for markup.
]]>
</employee>
```

CREATING VALID XML DOCUMENTS

What Are Valid XML Documents?

Definition: Well-formed XML validated against a DTD or schema is valid XML.

Definition: Both DTDs and schemas define the tags and attributes that are permissible within an XML document.

Besides checking to ensure that documents are well formed, parsers may also check to ensure that documents are valid. A "valid" XML document is a "well formed" XML document that also conforms to the rules of a document type definition (DTD) or XML schema. If a validating parser finds a validation error in an XML document, processing is halted.

In Java programming, classes are used as "templates" to create instances of objects. Think of a DTD or schema as a class and the conforming XML document as an object that is an instance of the "class" defined by the DTD or schema.

Why do we need to check for document validity? Isn't being well formed enough? In fact, valid XML is not necessary for many XML applications. The main reason for using a DTD or schema is to share data. If we wanted to share payroll data with another company, we would make sure our individual payroll.xml files referenced a common DTD to ensure that our accounting records followed the same format.

Valid XML Using a Document Type Definition (DTD)

Along with defining the tags and attributes that are permissible within an XML document, a DTD defines the document structure: what order tags should appear in and which tags can appear inside other tags. The DTD may also define entities we wish to use apart from the five entities predefined in the W3C XML spec. DTDs are specified within the W3C XML Recommendation (W3C, 2000).

A DTD can be declared within an XML document or as a reference to an external document. Normally we will reference an external DTD, as the whole point of using a DTD is to ensure that our data structure matches the structure of someone with whom we are sharing the data.

Simple XML Document with External DTD

Listing 2 shows our simple XML document modified to reference an external DTD document. Line 4 is the document type declaration used to reference the external file. The document type declaration must appear after an XML declaration but before the document's root tag. We will be taking a closer look at this declaration's syntax in the section on *Using XML to Share Data,* below.

Listing 2: File payroll2.xml—reference to external DTD.

```
 1 <?xml version="1.0" standalone="no"?>
 2 <!-- payroll2.xml -simple XML file
     that references external DTD -->
 3
 4 <!DOCTYPE payroll SYSTEM "C:\WINDOWS\
     Desktop\payroll.dtd">
 5
 6 <payroll company="Acme">
 7   <employee id="23">
 8      <first_name>John</first_name>
 9      <last_name>Doe</last_name>
10      <salary>$34,000</salary>
11   </employee>
12   <employee id="24">
13      <first_name>Jane</first_name>
14      <last_name>Smith</last_name>
15      <salary>$42,000</salary>
16   </employee>
17 </payroll>
```

Listing 3 shows the DTD referenced in payroll2.xml. External DTD file names use the ".dtd" extension. Note that payroll2.dtd is not a well-formed XML document. We will see later that schemas *are* well formed and valid XML and thus confer the benefits of parsability.

Listing 3: DTD for payroll2.xml.

```
 1 <?xml version="1.0"?>
 2 <!-- payroll2.dtd -the DTD for
     payroll2.xml -->
 3
 4 <!ELEMENT payroll (employee+)>
 5 <!ATTLIST payroll
 6   company CDATA #REQUIRED
 7 >
 8 <!ELEMENT employee (first_name,
     last_name, salary)>
 9 <!ATTLIST employee
10   id CDATA #REQUIRED
11 >
12 <!ELEMENT first_name (#PCDATA)>
13 <!ELEMENT last_name (#PCDATA)>
14 <!ELEMENT salary (#PCDATA)>
```

A DTD may include three types of declarations:

```
Element declaration:      <!ELEMENT>
Attribute declaration:    <!ATTLIST>
Entity declaration:       <!ENTITY>
```

For validity, every element, attribute, and nonstandard entity used in an XML document must be declared in its DTD.

An element declaration has the following syntax:

```
<!ELEMENT tagName elementContent>
```

As an example, look at Line 4 of Listing 3:

```
<!ELEMENT payroll (employee+)>
```

The tagName is `payroll`. The `payroll` root element must contain one or more `employee` elements (`employee+`).

An attribute declaration has the following syntax:

```
<!ATTLIST tagName attributeName valueType
  defaultValue>
```

As an example, look at Lines 5–7 of Listing 3:

```
<!ATTLIST payroll company CDATA #REQUIRED
```

The attributeName is `company`. As you can see in Listing 2, the `payroll` root element contains the `company` attribute, whose value is, in this case, "Acme." The "Acme" value is of type `CDATA`, or character data. The attribute does not have a default value but it is a `#REQUIRED` attribute.

An entity declaration has the following syntax:

```
<!ENTITY entityName "entityValue">
```

Listing 3 does not declare any entities, but here is an example of how they work:

DTD example:

```
<!ENTITY writer "Donald Duck.">
<!ENTITY copyright "Copyright W3Schools.">
```

XML example:

```
<author>&writer;&copyright;</author>
```

Simple XML Document with Internal DTD

Listing 4 shows the simple XML document, modified to include the DTD internally.

Listing 4: File payroll3.xml - DTD included internally.

```
<?xml version="1.0" standalone="yes"?>
<!-- payroll3.xml - simple XML file
  that includes internal DTD -->

<!DOCTYPE payroll [
   <!ELEMENT payroll (employee+)>
   <!ATTLIST payroll
     company CDATA #REQUIRED
   >
   <!ELEMENT employee (first_name,
     last_name, salary)>
   <!ATTLIST employee
     id CDATA #REQUIRED
   >
   <!ELEMENT first_name (#PCDATA)>
   <!ELEMENT last_name (#PCDATA)>
   <!ELEMENT salary (#PCDATA)>
]>

<payroll company="Acme">
  <employee id="23">
     <first_name>John</first_name>
     <last_name>Doe</last_name>
     <salary>$34,000</salary>
  </employee>
  <employee id="24">
     <first_name>Jane</first_name>
     <last_name>Smith</last_name>
     <salary>$42,000</salary>
  </employee>
</payroll>
```

Listing 4 has exactly the same functionality as the code shown in Listings 2 and 3. The only thing that is different is the syntax used to include the DTD internally.

Valid XML Using a Schema

XML schemas are an alternative to DTDs that are specified in a W3C Recommendation (World Wide Web Consortium (W3C), 2001). The schema specification addresses several shortcomings of DTDs (Lee, 2001), including the following:

A DTD is not a well-formed XML document, so XML parsers do not process it easily.

DTDs do not support data typing, so all content is treated as strings.

Schemas are well-formed XML documents that support data typing. The XML schema language is referred to as XML Schema Definition (XSD). Before the W3C's XSD was approved as a Recommendation, Microsoft implemented an interim solution to XSD—the XDR (XML-Data Reduced) schema language. XDR is supported in many Microsoft products such as the BizTalk server, Office 2000, Internet Explorer 5, and the SQL Server. Microsoft's parser, MSXML release 4, supports both the XSD and XDR languages. MSXML3, however, supports only XDR (Lee, 2001).

The purpose of an XML schema is to define the legal building blocks of an XML document, just like a DTD (Refsnes Data, 2002). An XML schema

defines elements that can appear in a document

defines attributes that can appear in a document

defines default and fixed values for elements and attributes

defines which elements are child elements and the order and number of child elements

defines whether an element is empty or can include text

defines data types for elements and attributes.

A detailed discussion of the XSD language is beyond the scope of this chapter, but we present below an example of an XSD schema and its XML document, along with a brief explanation.

Simple XML Document that References an External Schema

Listing 5 shows our simple XML document, which has been modified to reference an external XSD schema document, and Listing 6 shows the referenced schema file payroll4.xsd.

Listing 5: File payroll4.xml - reference to external Schema.

```
<?xml version="1.0" standalone="no"?>
<!-- payroll4.xml -simple XML file that
  references external Schema -->

<payroll company="Acme"
  xmlns:xsi="http://www.w3.org/2001/
    XMLSchema-instance"
  xsi:noNamespaceSchemaLocation="C:\
    WINDOWS\Desktop\payroll4.xsd">
    <employee id="23">
       <first_name>John</first_name>
       <last_name>Doe</last_name>
       <salary>$34,000</salary>
  </employee>
  <employee id="24">
     <first_name>Jane</first_name>
     <last_name>Smith</last_name>
     <salary>$42,000</salary>
  </employee>
</payroll>
```

Listing 6: Schema for payroll4.xml.

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- payroll4.xsd -the Schema for
  payroll4.xml -->

<xs:schema
 xmlns:xs="http://www.w3.org/2001/
   XMLSchema"
 elementFormDefault="qualified">

<xs:element name="payroll">
<xs:complexType>
```

```
10
11 <xs:sequence>
12       <xs:element name="employee"
           maxOccurs="unbounded">
13       <xs:complexType>
14
15       <xs:sequence>
16          <xs:element name="first_name">
17             <xs:simpleType>
18             <xs:restriction base=
                 "xs:string"/>
19             </xs:simpleType>
20          </xs:element>
21          <xs:element name="last_name">
22             <xs:simpleType>
23             <xs:restriction base=
                 "xs:string"/>
24             </xs:simpleType>
25          </xs:element>
26          <xs:element name="salary">
27             <xs:simpleType>
28             <xs:restriction base=
                 "xs:string"/>
29             </xs:simpleType>
30          </xs:element>
31       </xs:sequence>
32
33       <xs:attribute name="id"
           use="required">
34          <xs:simpleType>
35          <xs:restriction base=
              "xs:integer">
36             <xs:minInclusive value="0"/>
37             <xs:maxInclusive
                 value="100"/>
38          </xs:restriction>
39          </xs:simpleType>
40       </xs:attribute>
41
42       </xs:complexType>
43       </xs:element>
44 </xs:sequence>
45
46 <xs:attribute name="company"
     type="xs:string" use="required"/>
47
48 </xs:complexType>
49 </xs:element>
50
51 </xs:schema>
```

In Listing 5, the XSD reference takes the form of two attributes within the root element, as shown in lines 4–6. The XSD reference makes use of XML namespaces. We will be taking a closer look at XML namespaces in the section on *Using XML to Share Data*, below.

Listing 6 shows the XSD file referenced in payroll4.xml. External schema files use the ".xsd" extension. For validity, every element, attribute, and nonstandard entity used in an XML document must be declared in its schema.

Lines 4–6 compose the *root element*. The `<schema>` tag is the root element of every XML schema. Here, `<xs:schema>` specifies that the `"xs:"` prefix is to be used as a label for all succeeding tags in order to refer to the referenced namespace.

The remaining lines in the file specify elements and attributes to be used in payroll4.xml. The XML schema spec defines two types of elements: simple and complex. A simple element is an XML element that can contain only text. A complex element is an XML element that contains other elements and/or attributes.

PARSERS AND THEIR APIS

An XML parser is software that reads (and sometimes writes) an XML file and makes its data available for sharing, display, or further processing. In order to access parser methods and tell the parser what to do, we use an API (application programming interface).

XML parsers come in two flavors. *Nonvalidating parsers* only check that an XML document is well formed and do not check for validity against a DTD or schema. *Validating parsers* not only check for well-formedness but also verify that an XML document conforms to a DTD or schema.

APIs are available in most modern programming languages. It is possible to interface with parsers using Visual Basic, Java, ASP, Python, PHP4, Perl, JavaScript, etc. Of the available APIs, most implement methods that lcan process XML files in one or both of two ways: DOM (XML document object model) or SAX (simple API for XML).

Popular Parsers

The following validating parsers are the ones in most widespread use today:

Xerces—The Apache Software Foundation's open-source Java parser.

XML4J—IBM's XML Parser for Java. Free Java parser based on Xerces.

Oracle XML Parser—Available as a free standalone and also shipped with Oracle's RDBMS products, this parser supports Java, C, C++, and PL/SQL using industry-standard DOM and SAX interfaces.

XML::Parser—Larry Walls' perl-based parser based on James Clark's *expat*.

Microsoft Parsers—Microsoft has introduced several parsers, each more conformant to emerging XML standards than the last. Microsoft shipped Internet Explorer 4.0 with version 1.0 of the MSXML parser. Later products, including Office 2000, IE 5 and 5.5, Windows 2000, and BizTalk Server, were bundled with later versions of MSXML. Microsoft's current parser offerings include the following:

MSXML 3.0—A Component Object Model (COM) parser implemented via the file Msxml3.dll. Initial Web release: November 2000. Features complete implementation of XSLT and XPath W3C specs, improved SAX2, DOM, and namespace support.

MSXML 4.0 (Microsoft XML Core Services)—Another COM object-implemented parser (Msxml4.dll) that features DOM–SAX integration and complete support for XSD schemas. Initial Web release: October 2001.

System.Xml—The "assembly" that provides XML functionality for the .NET platform.

Many more parsers, both validating and nonvalidating, are available. Ken Sall has compiled an extensive list (Sall, 2000).

Which Parser Should We Use?

In practice, the parser we use will be determined by our development platform. If we plan to serve XML documents via a Perl-based CGI, then our choice will most likely be XML::Parser. For projects hosted by a Java application server, we will wind up with Xerces or XML4J or a similar Java parser. Reading and writing XML based on data in an Oracle database management system? Then stick with Oracle XML Parser. Building an end-to-end Microsoft solution? Use MSXML or System.Xml (Simpson, 2000).

DOM and SAX—The Standard API Models

Though APIs may implement many methods to facilitate parsing, two models have become standard: DOM (document object model) and SAX (simple API for XML). All current parsers support one or both of these models in their APIs.

DOM is a *tree-based API*. Tree-based APIs map XML documents into internal tree structures and then allow applications to navigate those trees. The Document Object Model (DOM) working group at the World-Wide Web Consortium (W3C) maintains a recommended tree-based API and there are many such APIs from other sources.

DOM's tree-like representation of XML documents are stored in memory, providing random access to the contents of an entire document. A DOM tree is composed of *nodes*. All nodes in the tree are contained within a root node that corresponds to the document's root element. All other elements in the XML document are also represented as nodes, as are data contained within the elements. Figure 2 shows how Microsoft's MSXML parser implements its built-in DOM API.

With DOM, much of the work of XML processing requires navigating the tree structure to find or modify data. DOM is useful when we need to create or modify a document in memory, read a document from an XML source file, or maintain ongoing random access to various parts of the XML document. The downside to DOM is that it puts a great strain on system memory, especially for large XML documents. For example, using DOM, a 100 kilobyte (KB) XML document can produce a tree that occupies up to 1 megabyte (MB) of memory.

SAX is an *event-based API*. Event-based APIs report parsing events (such as the start and end of elements) directly to the application rather than building internal trees. SAX was originally a Java-only API developed collaboratively by members of the XML-DEV mailing list, led by David Megginson. The API, now available in many languages, is supported by most parsers and has become a *de facto* API standard. The current version, SAX2, was released in May 2000.

SAX is designed for reading, not writing, XML documents. DOM is a better choice for modifying XML documents and saving changed documents to memory. SAX's advantages are that it uses much less memory than DOM and can locate specific pieces of data much faster than DOM.

SAX processes documents serially and generates events as it finds specific symbols in the XML document. In our application, we then implement handlers to deal with the events of interest to us, much as Java applications must handle events generated by a graphical user interface.

To understand how event-based processing works, consider this example. We have an XML document that contains stock prices:

```
<?xml version="1.0"?>
<stocks>
   <company>
       <symbol>IBM</symbol>
       <price>72.50</price>
   </company>
   <company>
      <symbol>CSCO</symbol>
      <price>13.05</price>
   </company>
</stocks>
```

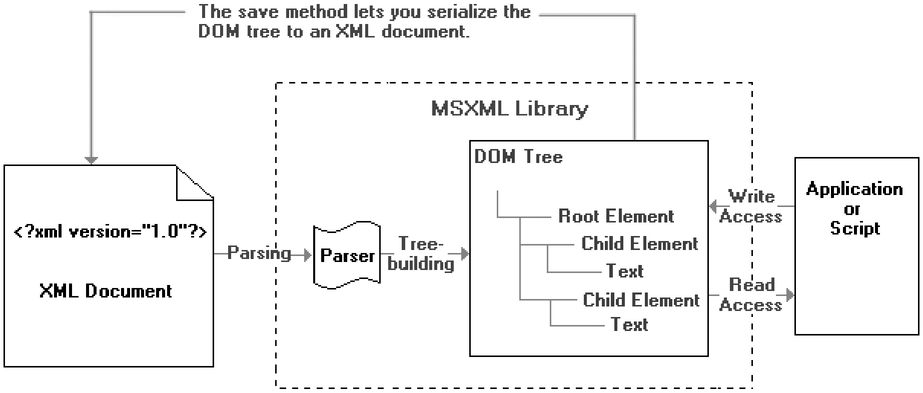


Figure 2: The DOM model. Source: Microsoft Corporation (2002a). © 2002 Microsoft Corporation. Reprinted with permission from Microsoft Corporation.

An event-based SAX interface would parse the above document and generate the following events:

```
start stocks
start element: stocks
start element: company
start element: symbol
characters: IBM
end element: symbol
start element: price
characters: 72.50
end element: price
end element: company
start element: company
start element: symbol
characters: CSCO
end element: symbol
start element: price
characters: 13.05
end element: price
end element: company
end element: stocks
end stocks
```

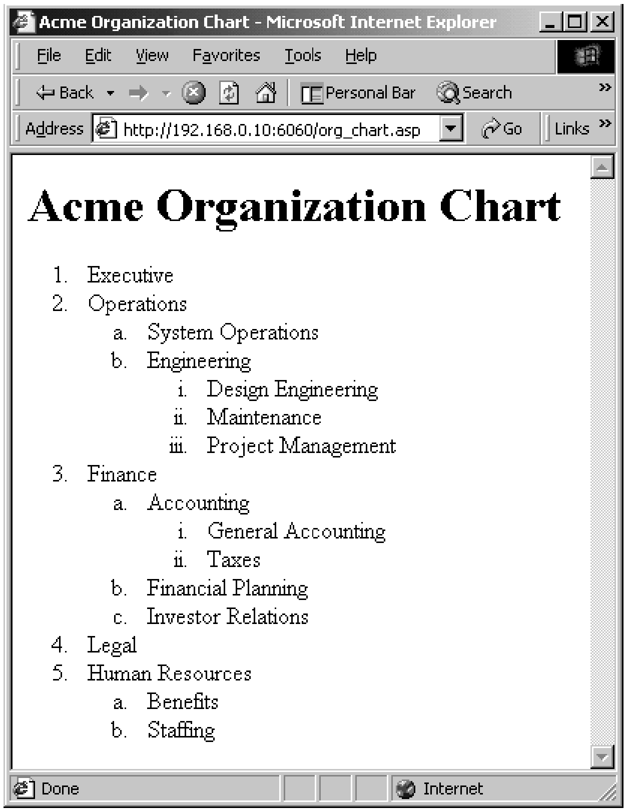
Now say that we are writing an application, and in our app we wish to locate the stock price of Cisco Systems (CSCO). Using the SAX interface, our app would have an easy time and be very efficient. The app's pseudocode might look something like this:

```
If event = "start element: symbol" Then
   If characters = "CSCO" Then
      For element: price
         Get characters
Stop parsing
```

A Parsing Example Using MSXML, ASP, and DOM

To get a feel for how parsing and APIs work, study the code below. This example uses Microsoft's MSXML parser running on a Windows 2000 server. IIS 5.0 processes the VBScript code in the ASP file. The VBScript code uses MSXML's built-in DOM API to traverse the XML document and dynamically generate HTML for display in a browser.

Our task is to generate an HTML page that shows Acme Corporation's organization chart. All organization data are contained in the file org_chart.xml, as shown in Listing 7. Listing 8, org_chart.asp, shows the VBScript code that uses DOM API to parse the XML document and generate HTML. Finally, Figure 3 shows the results of an HTTP request to the IIS Web server for the file org_chart.asp, as rendered in a browser.


Acme Organization Chart - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Personal Bar
Search
Address
http://192.168.0.10:6060/org_chart.asp
Go
Links
Done
Internet

Acme Organization Chart

1. Executive
2. Operations
   a. System Operations
   b. Engineering
      i. Design Engineering
      ii. Maintenance
      iii. Project Management
3. Finance
   a. Accounting
      i. General Accounting
      ii. Taxes
   b. Financial Planning
   c. Investor Relations
4. Legal
5. Human Resources
   a. Benefits
   b. Staffing

Figure 3: Acme Organization Chart as generated from org_chart.asp and org_chart.xml.

Listing 7: File org˙chart.xml.

```
<?xml version="1.0"?>
<!-- org_chart.xml -->

<Acme_Org_Chart>
  <department name="Executive"/>
  <department name="Operations">
    <department_sub1 name="System
      Operations"/>
    <department_sub1 name="Engineering">
        <department_sub2 name="Design
         Engineering"/>
        <department_sub2 name=
         "Maintenance"/>
        <department_sub2 name="Project
         Management"/>
   </department_sub1>
  </department>
  <department name="Finance">
   <department_sub1 name="Accounting">
       <department_sub2 name="General
        Accounting"/>
       <department_sub2 name="Taxes"/>
   </department_sub1>
   <department_sub1 name="Financial
     Planning"/>
   <department_sub1 name="Investor
     Relations"/>
  </department>
  <department name="Legal"/>
  <department name="Human Resources">
    <department_sub1 name="Benefits"/>
    <department_sub1 name="Staffing"/>
  </department>
</Acme_Org_Chart>
```

Listing 8: File org_chart.asp.

```
<% Option Explicit %>

<html><head><title>Acme Organization Chart</title></head>
<body>
<h1>Acme Organization Chart</h1>

<%
Dim xmlObject, fileString, xmlDoc

Set xmlObject = Server.CreateObject("Microsoft.XMLDOM") 'Fire up MS XML Parser
xmlObject.async = false
fileString = Server.MapPath("org_chart.xml") 'Map a path to the XML file
xmlObject.Load(fileString)
Set xmlDoc = xmlObject

Dim deptNodes, deptItems, deptSubItems
Dim deptName, deptItemName, deptSubItemName
Dim i, j, k

Set deptNodes = xmlDoc.selectNodes("//department")
Response.Write "<ol>" & vbCrLf

For i = 0 To (deptNodes.length - 1)
  deptName = deptNodes(i).getAttribute("name")
  Response.Write "<li>" & deptName & vbCrLf ' deptLevel0
  If deptNodes(i).hasChildNodes Then
     Response.Write "<ol type = 'a'>" & vbCrLf
     Set deptItems = deptNodes(i).selectNodes ("department_sub1")
     For j = 0 To (deptItems.length - 1)
         deptItemName = deptItems(j).getAttribute ("name")
         Response.Write "<li>" & deptItemName & vbCrLf
         If deptItems(j).hasChildNodes Then
              Response.Write "<ol type='i'>" & vbCrLf
              Set deptSubItems = deptItems(j).selectNodes("department_sub2")
              For k = 0 To (deptSubItems.length - 1)
                  deptSubItemName = deptSubItems(k).getAttribute("name")
                  Response.Write "<li>" & deptSubItemName & vbCrLf
              Next
              Response.Write "</ol>" & vbCrLf
              Set deptSubItems = Nothing
         End If
     Next
     Response.Write "</ol>" & vbCrLf
     Set deptItems = Nothing
  End If
Next

Response.Write "</ol>" & vbCrLf
Set deptNodes = Nothing
Set xmlDoc = Nothing
Set xmlObject = Nothing
%>

</body>
</html>
```

In Listing 8, Lines 8–14 start the parser, load the XML document, and build the DOM tree. In Line 11, the parser's `async` attribute is set to "false" to make sure that all the XML data are loaded before any other processing takes place. Lines 16–46 then call DOM API methods to traverse the tree and generate HTML for display in the requesting browser.

USING XML TO SHARE DATA

In order to communicate, computers need to agree on the rules that govern the exchange of information. For example, we store Web page information in HTML files but we need a set of rules—the hypertext transfer protocol (HTTP)—to transfer the page information from server to client.

So too, if we wish to share XML data between different platforms, applications, and companies, it is crucial that the different systems involved agree on the structure of data within the XML document. For example, if our computer always writes company name information to the third field of a record and another computer always looks for employee salary information in the third field, it's just not going to work. In order to accurately extract and use shared data, the computers have to agree on the type of information to be contained within each field.

This definition of data structure, familiar to database programmers, is known as a schema. To share information using XML, we do the same thing. We define a common data structure using either a DTD or an XML schema.

Once we agree on a common data structure, it does not matter which mechanism we use to exchange XML files. Our computer may request XML-structured data from another using a database query, RPC, DCOM, CORBA, or Web service. As long as the XML data on each end of the request meets validation criteria defined in a common DTD or schema, then our computer will have no problem using the data.

If it's just two people sharing data, then one might write a DTD and e-mail it to the other. The computers would then each reference a local copy of the agreed DTD to validate and exchange XML data. If all the companies in our industry wish to share data, then it's a bit more involved. We have to get all of the companies to agree on a common data structure for the exchange of information—an industry-standard DTD or schema.

Industry-Standard DTDs and Schemas

Those familiar with EDI (electronic data interchange) may be aware of the years of struggle required to hammer out standard industry guidelines. This issue remains the biggest challenge to using XML to exchange data. The rewards for coming to agreement are much lower cost, clearer implementation, and cross-platform interoperability. Progress has been made in several industries.

Early adopters of DTDs and schemas for XML data interchange include the financial services, energy/utilities, and healthcare sectors. To determine the state of DTD/schema development in a given sector, check with industry groups for the most current information in the field. Or have a look at the XML.org Industry Registry at http://www.xml.org/xml/registry.jsp, which maintains a repository of hundreds of industry-standard schemas and DTDs.

Referencing DTDs in Code

If we have agreed with our trading partner(s) on a DTD to use, then it is time to reference the DTD in our code. We do this by using the `<!DOCTYPE>` tag within our XML document (refer to the section on *Creating Valid XML Documents*, above). The `<!DOCTYPE>` tag may take one of the following four forms:

```
<!DOCTYPE rootElement SYSTEM "URI/URL">
<!DOCTYPE rootElement PUBLIC "URI/URL">
<!DOCTYPE rootElement PUBLIC
  "FormalPublicIdentifier">
<!DOCTYPE rootElement PUBLIC
  "FormalPublicIdentifier" "URI/URL">
```

In the first form, the `SYSTEM` keyword means that the DTD is stored locally on our system. As an example, consider the `<!DOCTYPE>` tag in List 2 in the *Creating Valid XML Documents* section:

```
<!DOCTYPE payroll SYSTEM "C:\WINDOWS\
  Desktop\payroll.dtd">
```

Here, the URL points to the path on the local file system where the payroll.dtd document resides.

The remaining three forms of the `<!DOCTYPE>` tag all use the PUBLIC keyword to indicate that the DTD is stored remotely somewhere on the Internet. We could modify the above `SYSTEM` example to access a publicly available version of the DTD as follows:

```
<!DOCTYPE payroll PUBLIC "http://www.
  web-feats.com/dtds/payroll.dtd">
```

The disadvantage of using a `PUBLIC` copy of the DTD is that looking up resources on the Internet is time-consuming and slows down the application every time the document is parsed. Once we agree on a standard DTD with our trading partner(s), it's common to copy the DTD and reference it via the local file system. The downside to this approach is that it ties the reference to a specific file, on a specific file system, in a specific location—if we move the application to a different server, we need to modify the DTD URL. Java programmers avoid this problem by packaging all required classes, XML files, and DTDs into a single JAR file and then including the JAR in their CLASSPATH.

Probably the best-known example of the third form, using a `PUBLIC DTD` with a "formal public identifier," is the following:

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0
  Transitional//EN">
```

This is the document type declaration used in an HTML 4.0 document. Here, `HTML` is the `rootElement;`

the root tag of every HTML document is `<HTML>`. `PUBLIC` means the DTD is available somewhere on the Internet. `"-//W3C//DTD HTML 4.0 Transitional//EN"` is the formal public identifier.

The formal public identifier (FPI) system is a carryover from SGML (simplified general markup language), the standardized language from which XML is derived. Public identifier fields are separated by `"//."` Here is what all the fields in the above FPI mean:

`"-//W3C//DTD HTML 4.0 Transitional//EN"`—the dash (`"-"`) indicates that the DTD is not a standard. Use a plus sign (`"+"`) if the DTD has been approved by a standards body such as ISO.

`"-//W3C//DTD HTML 4.0 Transitional//EN"`—W3C is the name of the owner who maintains the DTD.

`"-//W3C//DTD HTML 4.0 Transitional//EN"`—DTD is the document type; HTML 4.0 Transitional is a label that describes the document.

`"-//W3C//DTD HTML 4.0 Transitional//EN"`—two-letter abbreviation for the language of the documents to which the DTD applies. EN stands for English.

For our application to use a DTD referenced by an FPI, the FPI must be converted to an Internet-accessible URL. Our parser must have the ability to do this conversion. The parser must have what is known as an *entity resolver* in order to resolve the FPI into its actual URL.

The formal public identifier system has always been unwieldy and problem-laden. We are better off directly accessing `PUBLIC` DTDs by explicitly referencing their URLs, as in the previous document type declaration form. Better still, use a local copy of the DTD.

Finally, the fourth form of document type declaration uses both an FPI and a URL. For example:

```
<!DOCTYPE payroll PUBLIC "-//WebFeats//
  DTD Payroll//EN"
"http://www.web-feats.com/dtds/payroll.dtd">
```

Here, if the parser cannot resolve the URL of the DTD from the public identifier, it uses the URL as specified (Castro, 2001).

Referencing Schemas in Your Code

If we have agreed with our trading partner(s) on an XML schema to use, then we need to reference that schema in our code. We do this by adding attributes to the root element of our XML document.

If we are referencing a copy of the schema on the local file system, two attributes are added to the root element of our XML document. As an example, consider the root element in Listing 5 in the *Creating Valid XML Documents* section:

```
<payroll company="Acme"
   xmlns:xsi="http://www.w3.org/2001/
     XMLSchema-instance"
   xsi:noNamespaceSchemaLocation="C:\
     WINDOWS\Desktop\payroll4.xsd"
>
```

Here, `xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"` tells the validating parser that the `"xsi"` (XML schema instance) syntax we are using is defined in the W3C *namespace* `"http://www.w3.org/2001/XMLSchema-instance."` The parser does not actually access this W3C URL.

The `xsi:noNamespaceSchemaLocation="C:\WINDOWS\Desktop\payroll4.xsd"` fragment tells the parser to validate against an instance of the XML Schema `"payroll4.xsd"` that lives on the local file system along the path `"C:\WINDOWS\Desktop\payroll4.xsd."`

If we wished to access a publicly available version of the schema via the Internet from a *qualified XML namespace,* we would modify the above root element as follows:

```
<payroll company="Acme"
xmlns:xsi="http://www.w3.org/2001/
  XMLSchema-instance"
xmlns="http://www.web-feats.com/2002/
  Payroll"
xsi:schemaLocation="http://www.web-feats.
  com/2002/Payroll payroll4.xsd"
>
```

Again, `xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"` tells the parser that the syntax we are using is defined in the W3C namespace.

The fragment `xmlns="http://www.web-feats.com/2002/Payroll"` tells the parser the *qualified XML namespace* of the schema it will be using for validation; all of the elements used in the XML document are defined in the "http://www.web-feats.com/2002/Payroll" namespace.

Finally, `xsi:schemaLocation="http://www.web-feats.com/2002/Payroll payroll4.xsd"` is the Internet URL and file name that the parser has to access to fetch the payroll4.xsd that is to be validated against.

XML namespaces are defined in a W3C Recommendation (World Wide Web Consortium, 1999). They provide a means of avoiding element name conflicts when an XML document references additional, external XML documents. That is, the external XML schema that we are referencing is itself a well-formed XML document. Because we define your own tag names in XML, we need to ensure that if the schema uses any of the same tag names that we have defined, the parser can differentiate between the two different tags that have the same name. By prefixing tags from each document with prefixes for different namespaces, we avoid any potential conflicts.

Well-Formed, Valid with DTD, or Valid with Schema?—XML Design Considerations

When should we use a schema or DTD? When is well-formed XML sufficient?

In general, if we are not sharing our data with another application, platform, or company, then well-formed XML is sufficient. When our application's only goal is to use

XML to generate HTML for display in a browser, valid XML is usually not required. In fact, many browsers do not even have validating parsers. A DTD or schema adds a layer of processing between the point when a user clicks and when he or she sees the Web page. If there is no need to share data, we are better off avoiding the additional overhead of a DTD or schema.

When our application needs to share data with another party, the choice between using a DTD or schema comes down to these considerations:

Is there a standard DTD or schema available for our industry? If a standard already exists that is sufficient for our needs, then we should go with it. If a DTD is available, we should use the DTD. If an industry standard schema exists, we should use the schema.

If our organization only plans to use XML to share data internally, then consider these points in deciding between a DTD or schema:

Skill level of the programmers—XML schemas are more complex than DTDs.

Extensibility—Some day we may wish to share information outside of the organization. XML schemas are becoming the data definition language of choice as the standard has solidified.

Data types—If our application would benefit from set data types within the XML structure, then XML schema is a better choice. XML schema has support for data types, whereas DTDs view all data as strings.

Parsability—If our application would benefit from machine processing of the data structure, then XML schema is a better choice. XML schemas are well-formed XML documents and can be processed by parsers. DTDs are not well-formed XML.

USING XML IN THE BROWSER

There are five ways to view XML files in a browser:

Process XML on server and dynamically generate HTML for display in a browser.

Format an XML document for display using a cascading style sheet (CSS).

Format an XML document for display using XSL and/or XSLT.

Format an XML document for display using HTML and Microsoft data islands.

View XML source code or data structure.

Because of varying levels of XML implementation and conformance in different browsers and different browser versions, most XML processing today is done on the server. The server then dynamically generates HTML for display in the browser. We saw an example of this using VBScript, ASP, DOM, and MSXML in the section on Parsers and Their APIs, above. When processing XML on the server, we can programmatically control the generated HTML to target specific user agents (mobile devices or specific browser types) or specific locales (customized pages for French, German, English, etc., audiences).

When we have control over the browser to be used for viewing, as in developing for corporate intranets, it can be useful to have the browser itself directly format XML for display. In this section we will look at direct browser XML formatting using CSS and Microsoft data islands. XSL and XSLT are discussed in another chapter in this encyclopedia. We will also look at displaying XML source code in the browser. But first, we will look at XML-compatible browsers.

XML-Compatible Browsers

Newer browsers support parsing of XML documents to varying degrees. For the most part, XML-capable browsers do not check for XML validity but only for well-formedness. Some browsers, however, provide validating parsers and also may provide support for XSL and XSLT.

Although we generally use server-side processing to dynamically generate HTML for display in a browser, the following browsers are XML-capable:

Microsoft Browsers

Internet Explorer 4.x—Limited XML support.

Internet Explorer 5.x—Supports most of the international standards for XML 1.0, XSL and DOM.

Microsoft Internet Explorer 6.x—Conforms very well to the latest World Wide Web XSLT Recommendation.

Mozilla Browsers (parser based on James Clark's expat)

Netscape 6—Limited XML support.

Netscape 7 and Seamonkey 1.0—Support standards for XML 1.0 and DOM. Direct support for other XML languages, such as MathML (mathematical equations) and SVG (two dimensional vector diagrams).

Opera (versions 4.0 and higher)—XML support added May 2000.

Jumbo 3.0—A "molecular browser and toolkit." Jumbo was created to support the Chemical Markup Language (CML). Requires Java Runtime Environment 1.3.1 or later.

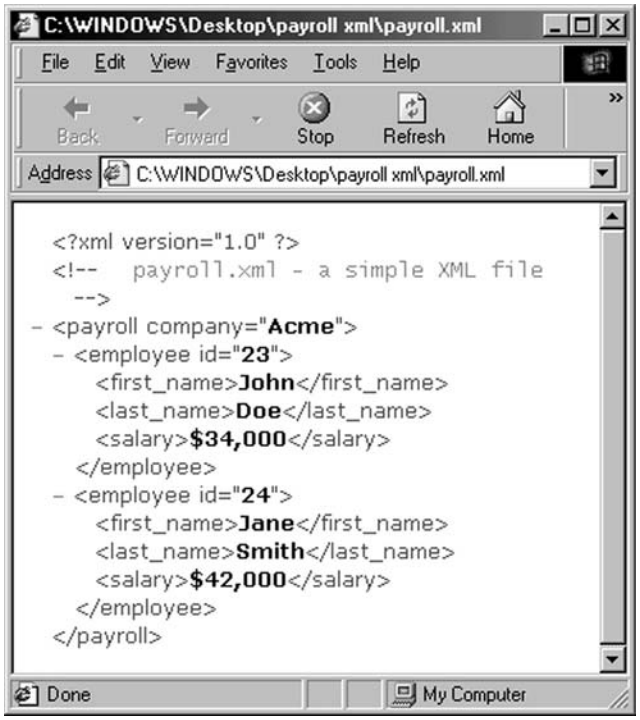
Viewing XML Source Code in a Browser

If we load an XML file into Internet Explorer, the browser renders it as a tree structure (see Figure 4). To view the XML source code, right-click on the page and select "View Source."

If we load an XML file into Seamonkey or Netscape 6 or 7, the browser parses the file and displays unformatted data (see Figure 5). To view the XML source code, right-click on the page and select "View Page Source."

Using Cascading Style Sheets to Format XML Data for Display in a Browser

Using cascading style sheets (CSS) to display XML data in a browser not only is dependent upon the browser's XML implementation; but also is subject to varying degrees of browser CSS support. Here we look at an


C:\WINDOWS\Desktop\payroll xml\payroll.xml
File Edit View Favorites Tools Help
Back
Forward
Stop
Refresh
Home
Address
C:\WINDOWS\Desktop\payroll xml\payroll.xml
Done
My Computer

```
  <?xml version="1.0" ?>
  <!--   payroll.xml - a simple XML file
    -->
- <payroll company="Acme">
  - <employee id="23">
      <first_name>John</first_name>
      <last_name>Doe</last_name>
      <salary>$34,000</salary>
    </employee>
  - <employee id="24">
      <first_name>Jane</first_name>
      <last_name>Smith</last_name>
      <salary>$42,000</salary>
    </employee>
  </payroll>
```

Figure 4: Internet Explorer 5.5 view of payroll.xml.

example of using CSS for XML display in Internet Explorer 5.5 and Netscape 7.

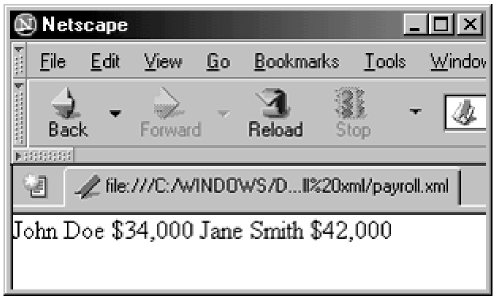
Listings 9 and 10 show an XML document that references an external style sheet and the style sheet itself, respectively.

In Listing 9, Line 4 is the processing instruction that links a style sheet to the XML document. It is understood by parsers in both Internet Explorer 5+ and Netscape 6+.

Listing 10 demonstrates *inheritance* in a cascading style sheet.

Listing 9: File payroll5.xml - reference to style sheet.

```
 1 <?xml version="1.0"?>
 2 <!-- payroll5.xml -simple XML file that
      references a CSS style sheet -->
 3
 4 <?xml-stylesheet type="text/css"
      href="style.css"?>
 5
 6 <payroll company="Acme">
 7   <employee id="23">
 8       <first_name>John</first_name>
 9       <last_name>Doe</last_name>
10       <salary>$34,000</salary>
11   </employee>
12   <employee id="24">
13       <first_name>Jane</first_name>
14       <last_name>Smith</last_name>
15       <salary>$42,000</salary>
16   </employee>
17 </payroll>
```


Netscape
File Edit View Go Bookmarks Tools Windov
Back
Forward
Reload
Stop
file:///C:/WINDOWS/D...ll%20xml/payroll.xml

```
John Doe $34,000 Jane Smith $42,000
```

Figure 5: Netscape Navigator 7 view of payroll.xml.

Listing 10: File style.css - style sheet used to format payroll5.xml in browser.

```
 1 /* style.css -style sheet used to
      format payroll5.xml in browser */
 2
 3 payroll {
 4   font-size: 16pt;
 5   font-weight: normal;
 6   font-family: Arial, Helvetica,
       Sans-Serif;
 7 }
 8
 9 employee {
10   display: block;
11 }
12
13 last_name {
14   font-weight: bold;
15 }
16
17 first_name {
18 }
19
20 salary {
21   position: absolute;
22   left: 200px;
23 }
```

Lines 3–7 set font properties for the entire document—because <payroll> is the root element of the XML file, all its child nodes inherit these settings.

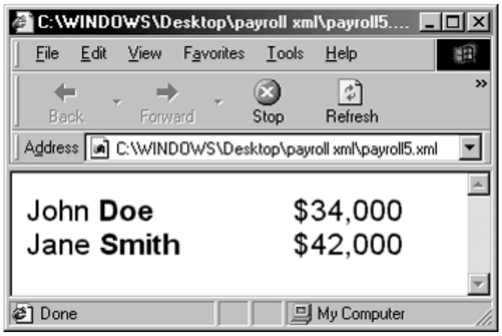
Lines 9–11 define each employee as a CSS *block*. This is necessary to display name and salary information on the same line; otherwise, each name and salary element would be treated as a separate block and each block would be rendered on a new line.

Lines 13–15 show how the last_name element's font-weight property overides the value inherited from the payroll element.

Lines 17–18 show that even though no properties are declared for the first_name element, font properties are inherited from the payroll element.

Lines 20–23 show use of absolute positioning to place the salary figures in their own column.

Figure 6 shows how payroll5.xml is rendered when loaded in IE 5.5. Figure 7 shows the file when loaded


C:\WINDOWS\Desktop\payroll xml\payroll5....
File Edit View Favorites Tools Help
Back Forward Stop Refresh
Address C:\WINDOWS\Desktop\payroll xml\payroll5.xml
Done
My Computer

John **Doe** $34,000
Jane **Smith** $42,000

Figure 6: Internet Explorer 5.5 view of payroll5.xml.

in Netscape 7—note the misformatting. This is due to Netscape's implementation of CSS absolute positioning.

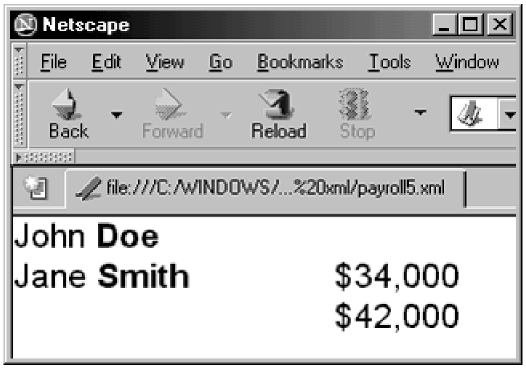
Using Microsoft Data Islands to Display XML Data in Internet Explorer

Internet Explorer lets us bind HTML tags to an XML data source included in the page as a "data island." Microsoft has proposed that data islands be included in the XML Recommendation, but so far W3C has not taken any action. Internet Explorer 5.0 and later are the only browsers that support data islands (Microsoft Corporation, 2002b).

Listing 11 shows an HTML file that uses data islands to parse and display data from the simple XML file "payroll.xml" (refer to Listing 1). IE identifies the `<xml>` tag as a data island. (There can be no <xml> tag within the referenced XML document.)

Listing 11: File payroll.htm.

```
 1 <html>
 2 <!-- payroll.htm -displays payroll.xml
     using Data Islands -->
 3
 4 <head>
 5 <title>Acme Payroll</title>
 6 <xml src="payroll.xml" id="xmldso"
     async="false"></xml>
 7 </head>
 8
 9 <body>
10 <h1>Acme Payroll</h1>
11
12 <table datasrc="#xmldso" width="300"
     border="1">
13   <thead>
14       <th>Last Name</th>
15       <th>First Name</th>
16       <th>Salary</th>
17   </thead>
18   <tr align="left">
19       <td><span datafld="last_name"
           style="font-weight: bold">
           </span></td>
20       <td><span datafld="first_name">
           </span></td>
21       <td><span datafld="salary">
           </span></td>
22   </tr>
23 </table>
24
25 </body>
26 </html>
```


Netscape
File Edit View Go Bookmarks Tools Window
Back Forward Reload Stop
file:///C:/WINDOWS/...%20xml/payroll5.xml

John **Doe**
Jane **Smith** $34,000
$42,000

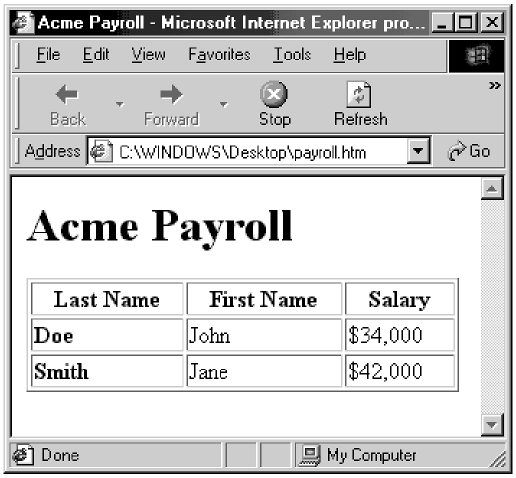
Figure 7: Netscape Navigator 7 view of payroll5.xml.

Line 6 uses the `<xml>` tag to load payroll.xml into an "invisible" Data Island called "xmldso." The `async="false"` attribute is added to the data island to make sure that all the XML data are loaded before any other HTML processing takes place.

Line 12 binds the XML data island to the `<table>` HTML element.

Lines 19–21 bind XML data fields to `<span>` elements inside the table data (`<td>`) HTML elements.

Figure 8 shows the file as rendered by IE.


Acme Payroll - Microsoft Internet Explorer pro...
File Edit View Favorites Tools Help
Back Forward Stop Refresh
Address C:\WINDOWS\Desktop\payroll.htm
Go
Done
My Computer

Acme Payroll

Last Name	First Name	Salary
Doe	John	$34,000
Smith	Jane	$42,000

Figure 8: Internet Explorer 5.5 view of payroll.htm.

USING XML CONFIGURATION FILES

A strong case could be made that configuration files are the most common use of XML today. They are certainly the most visible use of XML. Java application servers and .NET both depend on XML configuration files. Unix administrators are writing many of their config files in XML. Windows applications that previously used INI files or the Windows Registry now use XML configuration files.

Recent versions of RealNetworks RealAudio player use XML configuration files to store channels and tuner information—look at the channels.xml and tuner.xml files. In recent versions of Adobe Acrobat Readert the RdrENU.xml file is used to store version and update information.

The advantage of using XML configuration files over using Windows INI files, the Windows Registry, or Unix text files is that if we are writing an application for deployment on more than one platform, we can use the single XML config file everywhere rather than writing a different version of the config file for each platform.

Consider the following example, which stores an application's window position and most recently used file history (TurboPower Software Company, 2002):

```
<softapp>
  <mainwindow>
    <state size="normal" x="100" y="100">
    <MRUFile>
      <file>c:\bloat.txt</file>
      <file>d:\clients\wile\memo20010208.
        txt</file>
      <file>d:\rec\castlerock.txt</file>
      <file>c:\autoexec.bat</file>
    </MRUFile>
  </mainwindow>
</softapp>
```

Besides the widespread use of XML configuration files in Java application servers and .NET, more and more products are being built around XML. Table 2 shows some products that are noteworthy for their use of XML, either in configuration files or as an integral part of the product framework.

XML AND DATABASES

If we have XML, what do you need a database for? Isn't XML a programmatically accessible data source?

Bourret (2002a) answers,

> XML provides many of the things found in databases: storage (XML documents), schemas (DTDs, XML schema languages), query languages (XQuery, XPath, XQL, XML-QL, QUILT, etc.), programming interfaces (SAX, DOM, JDOM), and so on. On the minus side, it lacks many of the things found in real databases: efficient storage, indexes, security, transactions and data integrity, multiuser access, triggers, queries across multiple documents, and so on.

In real life, the need to use XML with databases arises for one of two reasons:

We need to access data from a "legacy" RDBMS (relational database management system); or

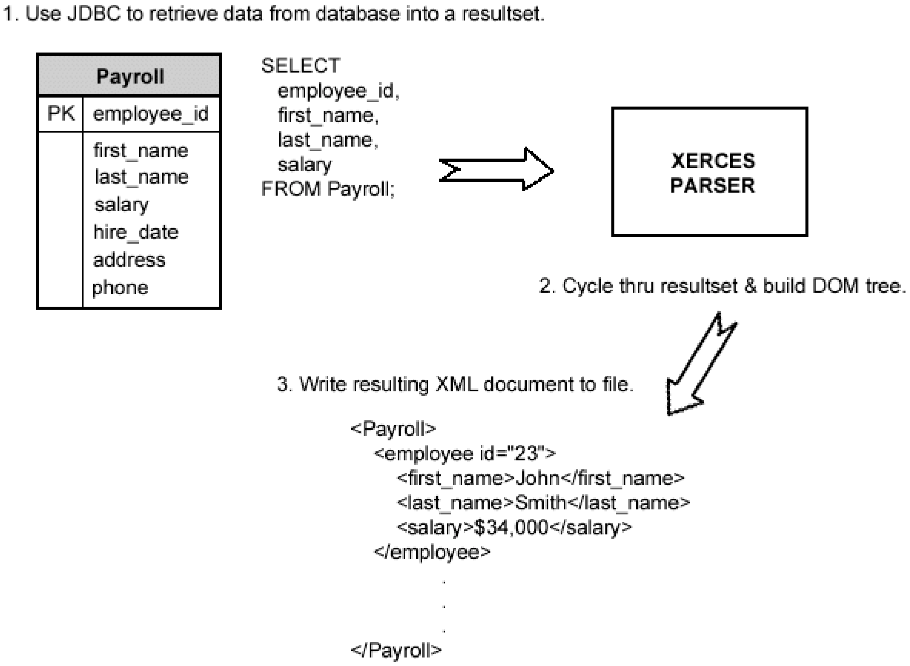
We need a repository for our XML documents; we need to "persist" XML data for long-term storage and retrieval.

To see where the first case may arise, consider this scenario. We are resellers of widgets. As orders come in from our customers, we need to order the widgets from our supplier. Orders come in to our company through our accounting software and are stored in a RDBMS. Our supplier, however, has provided us with a DTD and wishes to receive all our orders in XML format. So we must write an application that queries our RDBMS, extracts order information, and translates the orders to a valid XML format for sharing with our supplier. Our application is a *datacentric* use of XML—XML is used as a transport mechanism for data intended for machine consumption.

Contrast the preceding scenario with the following *document-centric* use of XML. Our Web site sells widgets to customers in England, France, and Germany. For each widget in our inventory, we have three XML documents, one for each of our target markets. England.xml contains a product description in English and a price in pounds. France.xml naturally contains the description in French

Table 2 Noteworthy XML Products

Product/Company	Description	URL
Ant/Apache Software Foundation	Ant is a Java-based build tool. All buildfiles generated by Ant take the form of standard XML documents	jakarta.apache.org/ant
Outlook 2002/Microsoft	Outlook 2002 lets you create and customize your own views. Each view is defined as an XML schema	msdn.microsoft.com/library/default.asp?url=/library/en-us/dnout2k2/html/odc_xmlviewdef.asp
Cocoon/Apache Software Foundation	Cocoon is an XML publishing framework that provides separation of content, logic and style	xml.apache.org/cocoon
BizTalk Server/Microsoft	BizTalk is an XML-based framework designed to facilitate electronic commerce over the Web	www.microsoft.com/biztalk


1. Use JDBC to retrieve data from database into a resultset.
Payroll
PK
employee_id
first_name
last_name
salary
hire_date
address
phone
XERCES PARSER
2. Cycle thru resultset & build DOM tree.
3. Write resulting XML document to file.

```
SELECT
  employee_id,
  first_name,
  last_name,
  salary
FROM Payroll;
```

```
<Payroll>
   <employee id="23">
      <first_name>John</first_name>
      <last_name>Smith</last_name>
      <salary>$34,000</salary>
   </employee>
         .
         .
         .
</Payroll>
```

Figure 9: SQL-to-XML using Java, Xerces, and Cloudscape.

and price in francs, whereas Germany.xml uses German marks. When a visitor to our Web site clicks a link for his or her preferred locale, he or she is rewarded by seeing all our glorious widgets described in his or her native tongue. So our Web application, upon user locale selection, must retrieve appropriate XML documents for each widget in our inventory. Our application then transforms XML data into HTML for viewing in a browser—document-centric XML is generally intended for human consumption.

The datacentric case utilizes a RDBMS. For the document-centric case, we may achieve best performance by using a NXD (native XML database).

XML AND RDBMS

In order to translate data between an RDBMS and XML, we must map the schema used in the RDBMS to the schema of our XML document. This is not a trivial task. Relational data mappings can be error-prone, inflexible, and time-consuming and, unless very carefully constructed, they can slow down our application's throughput.

To address the mapping problem, database vendors have developed tools that assist in converting XML documents to relational tables, and vice versa (Dayen, 2001).

Vendor Mapping Solutions

Oracle provides an XML SQL utility (XSU) that models XML document elements as a collection of nested tables. To go the other way, from SQL to XML, XSU constructs an XML document by using a one-to-one association between a table and a nested element.

IBM DB2 XML Extender provides new data types to store XML documents in DB2 databases and new functions to work with these structured documents. Managed by DB2, these documents are stored as character data or external files. Retrieval functions enable us to retrieve complete documents or individual elements.

Microsoft's SQL server extends SQL-92 with the FOR XML structure to provide SQL-to-XML mapping. For XML-to-SQL mapping, Microsoft has extended ADO by introducing the OPENXML row set.

An Example: SQL-to-XML Using Java, Xerces, and Cloudscape

Open source tools also provide methods for SQL-to-XML and XML-to-SQL mapping and data extraction. This example shows how to use Java JDBC classes to query a table in a Cloudscape database. The Xerces parser is invoked to construct a DOM tree that corresponds to the query result set. The Xerces serializer is then called to write the DOM tree to a flat file.

Figure 9 graphically depicts the process. The Payroll table in Acme's database is queried and Xerces builds the DOM tree and writes it to a file. Listing 12 shows the Java source code. Listing 13 shows the XML document that gets written to the flat file.

Listing 12: File SQLtoXML.java.

```
import java.sql.*;                 //JDBC Classes
import java.io.*;                  //Java IO classes for serialization
import org.w3c.dom.*;              //W3C DOM API classes
import org.apache.xerces.dom.*;    //Xerces DOM parser classes
import org.apache.xml.serialize.*; //Xerces classes for serialization

public class SQLtoXML {

//Declare database connection parameters
public static final String JDBC_URL =
  "jdbc:cloudscape:C:/Cloudscape_3.6/databases/Acme.db";
public static final String JDBC_DRIVER = "COM.cloudscape.core.JDBCDriver";

//Build query string
public static final String QUERY_STRING = "SELECT employee_id, first_name, "+
                                          "last_name, salary "+
                                          "FROM Payroll";

//Specify output file
public static final String OUTPUT_FILE = "c:\\payroll9.xml";

public static void main(String args[]) {

  try {

    //Connect to Cloudscape
    Class.forName(JDBC_DRIVER).newInstance();
    Connection conn = DriverManager.getConnection(JDBC_URL);

    //Query the database
    Statement statement = conn.createStatement();
    ResultSet payrollRS = statement.executeQuery(QUERY_STRING);

    //Call method to build DOM tree
    Document payrollXMLdoc = buildDOMtree(payrollRS);

    //Call method to serialize tree to file
    File outputFile = new File(OUTPUT_FILE);
    serializeDOM(payrollXMLdoc, outputFile);

    conn.close(); //Close Cloudscape connection
  } catch (Exception e) {
    System.out.println(e.toString());
  }
}

private static Document buildDOMtree(ResultSet_payrollRS) throws Exception {
  Document payrollXMLdoc = new DocumentImpl();

  //Create root element
  Element rootElement = payrollXMLdoc.createElement("Payroll");
  payrollXMLdoc.appendChild(rootElement);

  while (_payrollRS.next()) {

    //Build employee elements
    Element employee = payrollXMLdoc.createElement("employee");
    employee.setAttribute("id", _payrollRS.getString("employee_id"));

    //Create employee child elements
    Element first_name = payrollXMLdoc.createElement("first_name");
    Element last_name = payrollXMLdoc.createElement("last_name");
    Element salary = payrollXMLdoc.createElement("salary");

    //Populate employee children with data
    first_name.appendChild(payrollXMLdoc.createTextNode(
                _payrollRS.getString("first_name")));
    last_name.appendChild(payrollXMLdoc.createTextNode(
                _payrollRS.getString("last_name")));
    salary.appendChild(payrollXMLdoc.createTextNode(
                _payrollRS.getString("salary")));

    //Append employee children to employee
    employee.appendChild(first_name);
    employee.appendChild(last_name);
    employee.appendChild(salary);

    //Append employee to root element
    rootElement.appendChild(employee);
  }

  return payrollXMLdoc;
}

private static void serializeDOM(Document_payrollXMLdoc, File _outputFile)
                                        throws Exception {

  OutputFormat outputFormat = new OutputFormat("XML", "UTF-8", true);
  FileWriter fileWriter = new FileWriter(_outputFile);

  XMLSerializer xmlSerializer = new XMLSerializer(fileWriter, outputFormat);
  xmlSerializer.asDOMSerializer();

  xmlSerializer.serialize(_payrollXMLdoc.getDocumentElement());
}
}
```

Listing 13: Serialized XML document.

```
<?xml version="1.0" encoding="UTF-8"?>
<Payroll>
  <employee id="23">
    <first_name>John</first_name>
    <last_name>Doe</last_name>
    <salary>$34,000</salary>
  </employee>
  <employee id="24">
    <first_name>Jane</first_name>
    <last_name>Smith</last_name>
    <salary>$42,000</salary>
  </employee>
</Payroll>
```

XML AND NXD

Native XML databases (NXDs) are optimized for storage and retrieval of XML documents—documents go in and documents come out. Documents are the fundamental unit of storage, similarly to the way in which rows in a table are the fundamental unit of storage in a RDBMS. NXD's main benefit arises in document-centric XML usage—we do not have to worry about mapping our XML to some other data structure; we just insert data as XML and retrieve them as XML.

The NXD storage model is able to locate documents based on their content—elements, attributes, PCDATA, and element order. To facilitate location and retrieval of documents, NXDs support one or more *query languages*. XPath is the current NXD query language of choice but XQuery, a more database-oriented language, is currently under development. Several vendors have begun to release prototype XQuery implementations for use with their databases (Staken, 2001). XQL and a number of other languages are also used for querying XML documents.

To improve the performance of queries, NXDs support *document collections*. The notion of a document collection is similar to the concept of a table in a relational database. Where a RDBMS table is a collection of rows, an NXD document collection is a collection of XML documents that share common characteristics. For example, our company may store all XML documents related to sales orders in one collection, all XML documents related to product manuals in another collection, and so on. We can then query and manipulate a document collection as a set.

To do updates on documents, most NXDs currently require that we retrieve a document, change it using our favorite API, and then return it to the database. A few products, however, have built-in support for XML:DB XUpdate, an update language that facilitates easy, direct modifications to documents stored in a database.

Like other modern data management systems, most NXDs provide support for transactions, rollbacks, security, and concurrent access (Sholtz, 2002).

Table 3 A Sampling of Current NXD Products

Product	Company	License	URL
Ipedo XML Database	Ipedo	Commercial	www.ipedo.com
MindSuite XDB	Wired Minds	Commercial	xdb.wiredminds.com
Natix	data ex machina	Commercial	www.dataexmachina.de/natix.html
Oracle XML DB	Oracle	Commercial	otn.oracle.com/tech/xml/xmldb/content.html
Socrates XML	Cincom	Commercial	www.socratesxml.com/servlet/home.html
Tamino	Software AG	Commercial	www.softwareag.com/tamino
TEXTML Server	IXIA, Inc.	Commercial	www.ixiasoft.com/products/textmlserver
Virtuoso	OpenLink Software	Commercial	www.openlinksw.com/virtuoso
X-Hive/DB	X-Hive Corporation	Commercial	www.x-hive.com/products/db
ozone	Ozone	Open Source	ozone-db.org
Xindice	Apache Software Foundation	Open Source	xml.apache.org/xindice

NXD Vendors

Two of the more mature native XML databases products are Software AG's Tamino and Ipedo's Ipedo XML Database. Oracle's XML DB is a feature of Oracle9*i* Database and provides native XML storage and retrieval technology. The Apache Software Foundation's Xindice is a promising open source NXD.

Table 3 lists several vendors' NXDs. Many more NXDs are available. Ronald Bourret maintains an extensive list (Bourret, 2002b).

Using NXDs

Writing code to store, query, and retrieve documents from a NXD depends on the NXD product and the API methods it exposes. Although NXD code samples are beyond the scope of this chapter, many excellent resources appear on the Web, including the following:

XML Querying: XPath and XQuery (Obasanjo, 2001)—Good introduction to the two leading NXD query languages.

Apache Xindice (Apache Software Foundation, 2002)—Download the latest release. Untar/unzip and refer to the /java/examples directory for examples using Java, XPath, and the *XML:DB API*.

XML Tech Center—Sample Corner (Oracle, 2002)—Oracle-centric code for using both Oracle's NXD-like product (Oracle9*i* XML) and their standard RDBMS product.

USEFUL TOOLS

Any text editor is sufficient for writing XML. Many vendors, however, have created custom XML editing tools. These tools can be a real productivity enhancer, as many of them incorporate some or all of the following features:

XML editing and validation

Schema editing and validation

XSL editing and transformation

Automatic DTD generation from XML document and vice versa

Automatic Schema generation from XML document and vice versa

IDE workspace with multiple document views and context-sensitive property inspectors.

Table 4 shows a few of the more popular XML editors. Many more authoring tools are available. Ken Sall maintains an extensive list (Sall, 2002).

THE FUTURE

What lies in the future for XML? Start studying now; one of the topics below is sure to become the Next Big Thing.

Web Services

Web services are an XML-based means of invoking methods and programs across a distributed computing environment. We are talking about application-to-application communication. Previously this role was filled by such diverse communications protocols as COM, CORBA, DCOM, and RMI. The problem with these protocols is that they are not easily interoperable—COM can only call COM, CORBA can only call CORBA. It is not easy for an application running on a Windows desktop to call an application running on an IBM mainframe.

Web services solve this problem. Surprise, a standard that all competing camps agree upon! Think of Web

Table 4 XML Editors

Product	Company	License	URL
XML Spy	Altova	Commercial	www.xmlspy.com
XMetaL	Corel/SoftQuad	Commercial	www.xmetal.com
xmlPro	Vervet Logic	Commercial	www.vervet.com
XMLwriter	Wattle Software	Commercial	www.xmlwriter.net
XML Notepad	Microsoft	Freeware	msdn.microsoft.com/xml/notepad/intro.asp

services as RMI with a marketing department. Web services let developers invoke services running on local and remote resources using a platform-independent range of client applications.

Consider this example. An insurance company has agents preparing quotations from headquarters, branch offices, and remote dialup connections. To prepare the quotations, software being run by the agents needs to access actuarial tables. If the tables are included within each locally installed software package, then every time the tables change, all 6,000 agents in the company must be sent CD-ROMs with the new data. The Web services solution? Host the actuarial tables (the business rules) at one central location and have all the distributed applications request data from the central repository. That way, when the rules change, all clients immediately have access to the latest data. The general principles are (a) encapsulate the business rules, (b) provide methods to access them, and (c) make the methods accessible over the intranet or Internet.

The Web services solution is implemented using three XML-based languages:

WSDL (Web services definition language)
SOAP (simple object access protocol)
UDDI (Universal description, discovery, and integration).

A WSDL document exposes the methods of the application being offered as a Web service. The location (URI) of the WSDL is registered in a UDDI repository and accessed via a UDDI request. SOAP provides the messaging protocol used in the client-to-UDDI-to-WSDL-to-service communication.

Several companies already offer free usage of public UDDI repositories. For example, IBM maintains a registry at https://uddi.ibm.com/ubr/registry.html, Microsoft at http://uddi.microsoft.com/default.aspx, and Oracle at http://otn.oracle.com/tech/webservices/htdocs/uddi/content.html.

It is not necessary to use a public UDDI repository. Tools are readily available that let companies create their own private UDDI registries hosted on the company intranets. In fact, this author, for one, sees intranets as the primary consumer of Web services. Web services that are offered over the public Internet suffer from the Scylla and Charybdis of latency and security concerns.

ebXML

ebXML (electronic business extensible markup language) is a joint initiative of the United Nations (UN/CEFACT) and OASIS, whose goal is to create a single global electronic market.

To facilitate global commerce, the ebXML business process specification provides a standard framework for defining business transactions. The current version of ebXML supports binary collaboration (business between two parties). Future versions will support business collaboration among multiple parties (Siddalingalah, 2001).

ebXML draws upon ideas similar to Web services. To interact with a trading partner, a company first accesses an ebXML registry to determine business rules both parties can agree on to govern the transaction. Current proposals call for using UDDI and SOAP to access the ebXML registry.

ebXML is not so much a new specification as a codification of existing standards and business practices. Whereas ebXML hopes to succeed EDI (electronic data interchange), official descriptions tend to emphasize learning from EDI rather than throwing it out. The ebXML specifications tend to incorporate what businesses do anyway, rather than trying to get companies to do business differently. The ebXML initiative clearly holds an embrace-existing-standards-and-methods attitude (Mertz, 2001).

VoiceXML

VoiceXML is a server-side technology for creating voice interfaces. It uses speech recognition or touchtone keypads for input and prerecorded audio or text-to-speech synthesis for output.

One popular application of VoiceXML is the voice portal, a telephone service where callers dial a phone number to retrieve information such as stock quotes, sports scores, or weather reports. Other possible applications include voice-enabled intranets, contact centers and notification services, and more exotic applications such as speech-controlled home appliances (Rehor, 2001).

Although voice-based technologies are likely to see continued development, conflicting standards may impede the acceptance of VoiceXML. VoxML, introduced by Motorola in September 1998, is a more mature technology than VoiceXML but contains less advanced features. Microsoft is active in SALT, another voice XML technology introduced in 2002.

Because “voice browser” technology requires extensive server-side architecture, companies that have invested in their own implementation schemes may be unlikely to abandon them.

GLOSSARY

Data-centric Type of application that uses XML as a transport mechanism for data intended for machine consumption.

Dialect Any XML-based language that conforms to W3C’s syntactic and semantic rules. Usually designed to address one particular class of data, such as musical or chemical notation. Also known as a *grammar*.

Document-centric Type of application characterized by the display of XML data for human consumption.

DOM (Document Object Model) An API (Application Programming Interface) that maps an XML document into an internal tree structure.

DTD (document type definition) Used to check the *validity* of an XML document; defines the order tags should appear in and which tags can appear inside what other tags. May also define *entities*.

Element The basic unit of data storage in an XML document. An element consists of a start tag, an end tag, and everything between the start and end tags.

Entity An “escape sequence” that starts with an ampersand (“&”) and ends with a semicolon (“;”). Entities provide a shorthand notation for special characters, words, or phrases that may be used within an XML document.

Grammar See dialect.

Namespace A means to avoid element name conflicts when more than one XML document is referenced within a single application.

Parser The application that reads and interprets XML. Also known as an *XML processor*. Parsers reject XML documents that are not *well-formed* and may also check for *validity*.

SAX (simple API for XML) An API (application programming interface) that reports parsing events (such as the start and end of elements) directly to the application rather than building an internal tree.

Schema Like a DTD, a schema is used to check the *validity* of an XML document. Unlike DTDs, schemas are well-formed XML documents that support data typing.

Valid Well-formed XML that is validated against a DTD or schema is valid XML.

Well-formed XML with correct syntax is well-formed XML.

XML (extensible markup language) A text-based, human- and machine-readable language used for sharing data over Web, intranets, and extranets.

XML processor See parser.

CROSS REFERENCES

See *Databases on the Web; Extensible Stylesheet Language (XSL); HTML/XHTML (HyperText Markup Language/Extensible HyperText Markup Language); Web Services*.

REFERENCES

Apache Software Foundation (2002). Apache Xindice. Retrieved September 28, 2002, from http://xml.apache.org/xindice

Bourret, R. (2002a). XML and Databases. Retrieved September 24, 2002, from http://www.rpbourret.com/xml/XMLAndDatabases.htm

Bourret, R. (2002b). XML database products: Native XML databases. Retrieved September 24, 2002, from http://www.rpbourret.com/xml/ProdsNative.htm

Castro, E. (2001). *XML for the World Wide Web Visual Quickstart Guide*. Berkeley, CA: Peachpit Press.

Dayen, I. (2001). Storing XML in relational databases. Retrieved September 19, 2002, from http://www.xml.com/pub/a/2001/06/20/databases.html

Holzner, S. (2001). *Inside XML*. Indianapolis, IN: New Riders Publishing.

Lee, W. M. (2001, October/November). Applying XML schema to XML documents. *XML Magazine,* 30–38. Retrieved September 6, 2002, from http://www.fawcette.com/xmlmag/2001_11/magazine/columns/collaboration/wlee/

Mertz, D. (2001). Understanding ebXML. Retrieved October 6, 2002, from http://www-106.ibm.com/developerworks/xml/library/x-ebxml

Microsoft Corporation (2002a). Microsoft XML core services (MSXML) 4.0—DOM developer's guide. Retrieved October 10, 2002, from http://msdn.microsoft.com/library/default.asp?url = /library/en-us/xmlsdk/htm/dom_concepts_2lkd.asp

Microsoft Corporation (2002b). Microsoft XML core services (MSXML) 4.0—XML developer's guide. Retrieved October 3, 2002, from http://msdn.microsoft.com/library/default.asp?url = /library/en-us/xmlsdk/htm/xml_concepts2_2n03.asp

Obasanjo, D. (2001). An exploration of XML in database management systems. Retrieved September 29, 2002, from http://www.25hoursaday.com/StoringAndQueryingXML.html

Oracle Corporation (2002). XML tech center—sample corner. Retrieved September 29, 2002, from http://technet.oracle.com/sample_code/tech/java/xml/content.html

Refsnes Data (2002). XML schema tutorial. Retrieved September 7, 2002, from http://www.w3schools.com/schema/

Rehor, K. G. (2001). What is VoiceXML? Retrieved October 6, 2002, from http://www.voicexmlreview.org/Jan2001/features/Jan2001_what_is_voicexml.html

Sall, K. (2000). XML software guide: XML parsers. Retrieved August 16, 2002, from http://wdvl.internet.com/Software/XML/parsers.html

Sall, K. (2002). XML software guide: XML and XSL editors. Retrieved August 16, 2002, from http://wdvl.internet.com/Software/XML/editors.html

Sholtz, P. (2002, October). Tame the information tangle. *New Architect Magazine,* 7(10), 36–40. Retrieved September 28, 2002, from http://www.newarchitectmag.com/documents/s = 2453/na1002d/index.html

Siddalingalah, M. (2001). Overview of ebXML. Retrieved September 30, 2002, from http://dcb.sun.com/practices/webservices/overviews/overview_ebxml.jsp

Simpson, J. E. (2000). Choosing an XML parser. Retrieved September 8, 2002, from http://www.xml.com/pub/a/2000/08/23/whichparser/index.html

Staken, K. (2001). Introduction to native XML databases. Retrieved September 28, 2002, from http://www.xml.com/pub/a/2001/10/31/nativexmldb.html

TurboPower Software Company, 2002. Using XML in your next project. Retrieved October 2, 2002, from http://www.turbopower.com/products/xmlpartner/using/

World Wide Web Consortium (W3C) (1999). Namespaces in XML. "Recommendation." Retrieved August 15, 2002, from http://www.w3.org/TR/REC-xml-names

World Wide Web Consortium (W3C) (2000). Extensible Markup Language (XML) 1.0. "Recommendation." Retrieved August 13, 2002, from http://www.w3.org/TR/2000/REC-xml-20001006

World Wide Web Consortium (W3C) (2001)2 XML Schema. "Recommendation." Retrieved August 15, 2002, from http://www.w3.org/TR/xmlschema-0/

FURTHER READING

Jones, J. (2002). XML 101. Retrieved August 13, 2002, from http://www.swynk.com/friends/jones/articles/xml_101.asp

Martin, D., and 12 other authors (2000). *Professional XML*. Birmingham, UK: Wrox Press.

Extensible Stylesheet Language (XSL)

Jesse M. Heines, *University of Massachusetts Lowell*

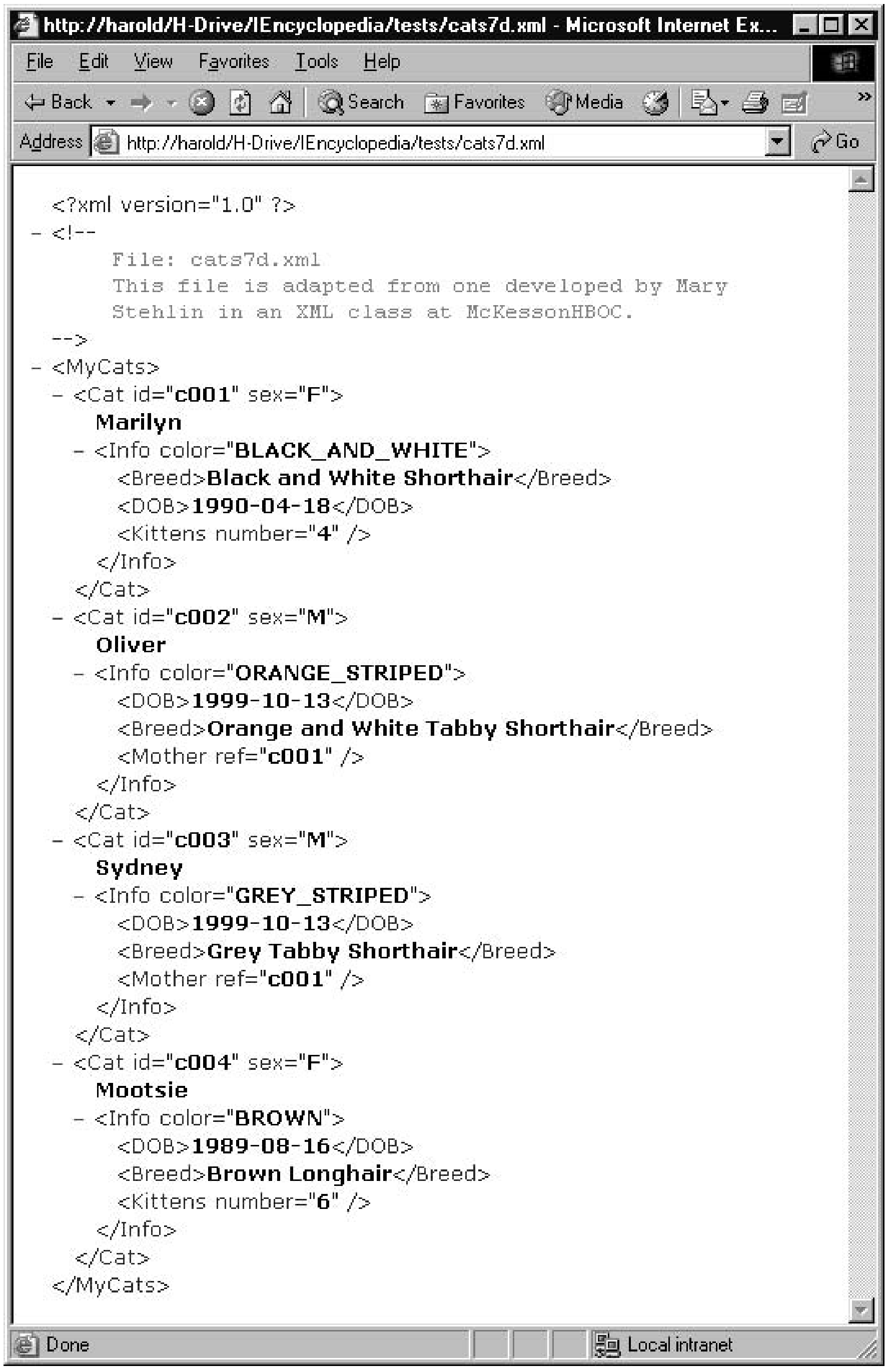
XML, XML EVERYWHERE, BUT NOT A DROP OF COMPATIBILITY

If we both store our data as XML, we should be compatible with each other, right? Unfortunately, no. The XML Recommendation specifies how XML documents are formed, but not how the data they contain should be organized. Consider, for example, the many valid ways to store a date:

```
<date>February 26, 2002</date>  <!-- American standard     -->
<date>26 February 2002</date>   <!--  European standard    -->
<date month="February" day="26" year="2002"/>  <!-- no ambiguity -->
<date>                          <!-- reasonable alternative   -->
   <month>Feb</month>             <!--  which also eliminates  -->
   <day>26</day>                  <!--  ambiguity but now uses -->
   <year>2002</year>              <!--  an abbreviation for the  -->
</date>                         <!-- month name             -->
<date>2/26/02</date>            <!-- common American abbreviation -->
<date>26.2.2002</date>          <!-- common European abbreviation -->
<date>2002-02-26</date>         <!-- ISO 8601 format, standard used -->
                                <!--    in XML Schemas          -->
<date>37313</date>              <!-- serial number  -->
```

The simple solution, of course, is for everyone to agree to express dates in the same format. As the flower girl cum socialite mused in *My Fair Lady,* "Wouldn't it be loverly?" But as the auctioneer cum Mafioso chirped in *Mickey Blue Eyes,* "Fughedaboudit!"

The problem is not that people disdain compatibility, it's that XML allows each of us to wrap our data in any tags that make sense to us alone. There are no standards for tag names. So even though Company A may really want its inventory system to automatically communicate with Company B's order processing system, and even though Company A's sending program can generate XML output and Company B's receiving program can take XML input, the two may still not be able to communicate. Both may be Y2K compliant, but if Company A's program represents the year number in an element (`<year>2002</year>`) and Company B's program expects an attribute on a `date` tag (`year="2002"`), well, "never the twain shall meet." The analogy here is that if you ask for a "hoagie" in a town where such sandwiches are known as "grinders," "heroes," or "subs," you'll very likely go hungry. "But this is no big deal," you say. "All one has to do is write a simple C++ or Java program to convert one data format to another." I fear that such programs are not as simple as they may at first appear, and of course the more complex the conversion, the more complex the conversion program. XML data are fairly easy to create and maintain, while C++ and Java programs are not. It would be nice if we had an


http://harold/H-Drive/IEncyclopedia/tests/cats7d.xml - Microsoft Internet Ex...
File Edit View Favorites Tools Help
Back
Search
Favorites
Media
Address
http://harold/H-Drive/IEncyclopedia/tests/cats7d.xml
Go
Done
Local intranet

```
  <?xml version="1.0" ?>
- <!--
     File: cats7d.xml
     This file is adapted from one developed by Mary
     Stehlin in an XML class at McKessonHBOC.
  -->
- <MyCats>
  - <Cat id="c001" sex="F">
      Marilyn
    - <Info color="BLACK_AND_WHITE">
        <Breed>Black and White Shorthair</Breed>
        <DOB>1990-04-18</DOB>
        <Kittens number="4" />
      </Info>
    </Cat>
  - <Cat id="c002" sex="M">
      Oliver
    - <Info color="ORANGE_STRIPED">
        <DOB>1999-10-13</DOB>
        <Breed>Orange and White Tabby Shorthair</Breed>
        <Mother ref="c001" />
      </Info>
    </Cat>
  - <Cat id="c003" sex="M">
      Sydney
    - <Info color="GREY_STRIPED">
        <DOB>1999-10-13</DOB>
        <Breed>Grey Tabby Shorthair</Breed>
        <Mother ref="c001" />
      </Info>
    </Cat>
  - <Cat id="c004" sex="F">
      Mootsie
    - <Info color="BROWN">
        <DOB>1989-08-16</DOB>
        <Breed>Brown Longhair</Breed>
        <Kittens number="6" />
      </Info>
    </Cat>
  </MyCats>
```

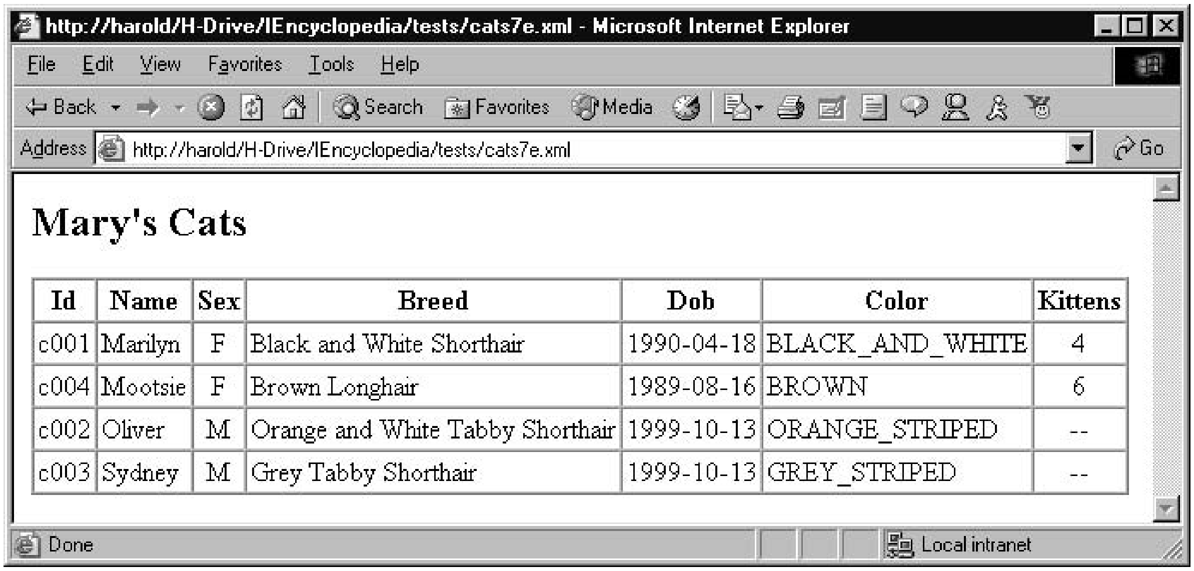
Figure 1: The Internet Explorer default rendition of XML data. (This file is adapted from one developed by Mary Stehlin in an XML class taught by the author at McKessonHBOC, Inc., in Alpharetta, GA, May 2000.)

XML-based solution to convert XML data from one format to another. This is precisely what XML stylesheet language *transformation* (XSLT) is for.

A "stylesheet language" may seem a strange moniker for a protocol that converts one data format to another, but the name is historical. XSLT is a component of the extensible stylesheet language (XSL), which was originally conceived as a native XML replacement for cascading style sheets (CSS), just as the schema was conceived as a native XML replacement for document type definitions (DTDs). The XSLT part of XSL is an extremely powerful and efficient method for viewing XML data in a variety of formats. It does this by essentially transforming XML into HTML. A few examples will make this readily apparent.

Formatting XML-based Web Pages

If one brings up XML data in Internet Explorer, one sees a nice tree-structured rendition such as that in Figure 1. Microsoft has cleverly made this display interactive,


http://harold/H-Drive/IEncyclopedia/tests/cats7e.xml - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://harold/H-Drive/IEncyclopedia/tests/cats7e.xml
Go
Done
Local intranet

Mary's Cats

Id	Name	Sex	Breed	Dob	Color	Kittens
c001	Marilyn	F	Black and White Shorthair	1990-04-18	BLACK_AND_WHITE	4
c004	Mootsie	F	Brown Longhair	1989-08-16	BROWN	6
c002	Oliver	M	Orange and White Tabby Shorthair	1999-10-13	ORANGE_STRIPED	--
c003	Sydney	M	Grey Tabby Shorthair	1999-10-13	GREY_STRIPED	--

Figure 2: The same data formatted with XSL.

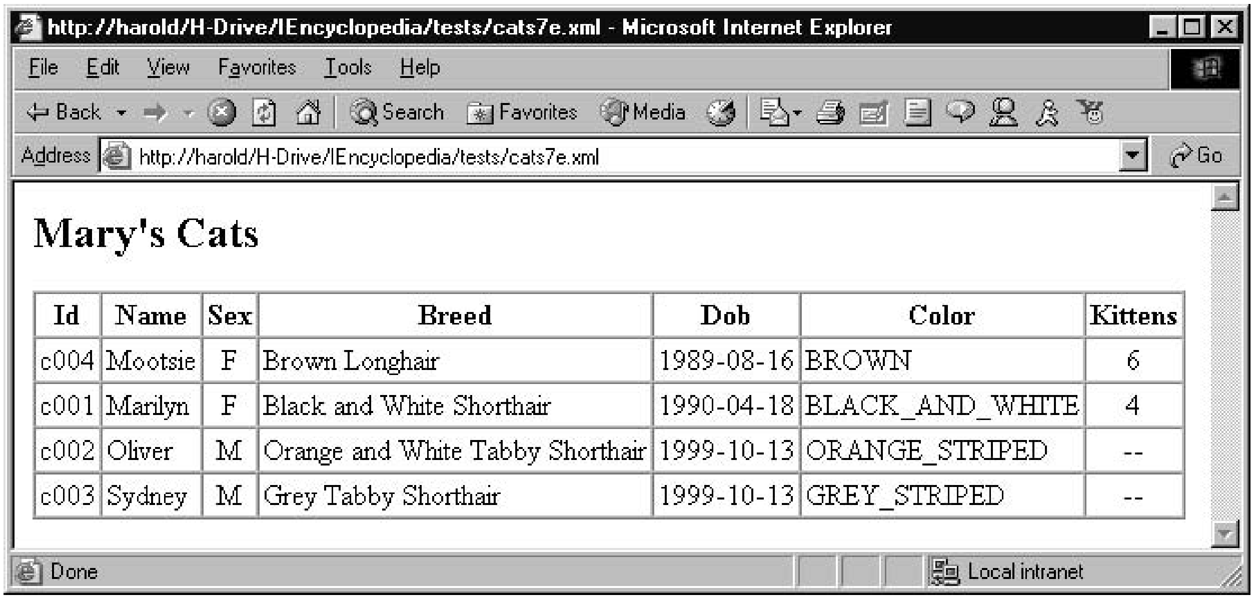
allowing the user to expand and contract subtrees by clicking the + and − prefixes. This is nice for developers, but rather awkward for regular folks who are interested in the data per se, not its XML representation.

A relatively few lines of XSL can transform these data into the user-friendly format shown in Figure 2. In this example, XSL was used to generate the entire HTML page being displayed, but one could just as easily use XSL to format just part of the page. Thus dynamic data stored in XML format can be integrated with dynamic or static data stored in other Web-friendly formats to produce pages that not only look good, but also allow more sophisticated processing because the data have semantics (expressed in XML tags) rather than just formatting information (expressed in HTML tags).

(For readers who want to dive right into code, listings for this and the examples in Figures 3 and 4 are provided in the Appendix. Also note that the Internet Explorer XSL Engine will not apply XSL formatting to an XML file that contains a schema reference on the root node!)

Generating Reports from XML Data

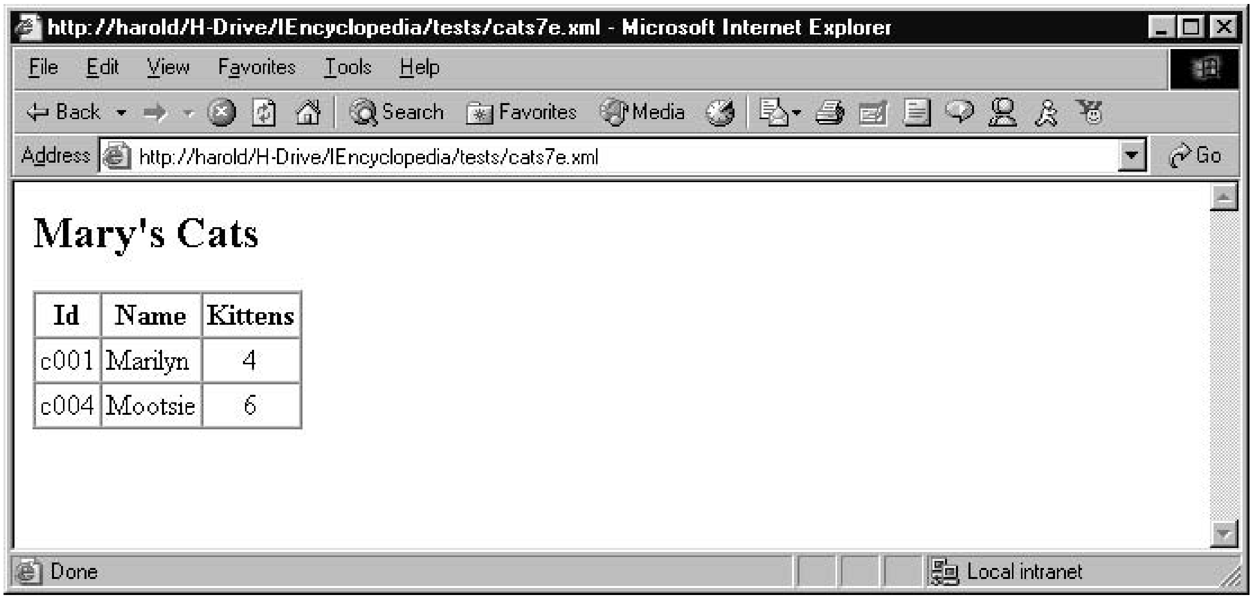
If one looks carefully at Figures 1 and 2, one will notice that "Mootsie" appears as the fourth cat in Figure 1, yet is listed second in Figure 2. This is an indication that XSL can do more than just format data: It can also rearrange it. Rearrangement is another simple type of transformation (Cagle, 2000a). XSL can also *filter* XML data, that is, extract selected elements whose values match a given *pattern*


http://harold/H-Drive/IEncyclopedia/tests/cats7e.xml - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://harold/H-Drive/IEncyclopedia/tests/cats7e.xml
Go
Done
Local intranet

Mary's Cats

Id	Name	Sex	Breed	Dob	Color	Kittens
c004	Mootsie	F	Brown Longhair	1989-08-16	BROWN	6
c001	Marilyn	F	Black and White Shorthair	1990-04-18	BLACK_AND_WHITE	4
c002	Oliver	M	Orange and White Tabby Shorthair	1999-10-13	ORANGE_STRIPED	--
c003	Sydney	M	Grey Tabby Shorthair	1999-10-13	GREY_STRIPED	--

Figure 3: Cat data sorted by date of birth (DOB).


http://harold/H-Drive/IEncyclopedia/tests/cats7e.xml - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://harold/H-Drive/IEncyclopedia/tests/cats7e.xml
Go
Done
Local intranet

Mary's Cats

Id	Name	Kittens
c001	Marilyn	4
c004	Mootsie	6

Figure 4: Cat data filtered to show only selected data in selected elements of the two female cats.

and, if so desired, juxtapose their data values with those of other selected elements. Such “filtering” is similar to “selecting” data from relational databases with SQL queries that incorporate `WHERE` clauses. Furthermore, XSL can *sort* data on the values of any XML element, allowing data to be presented in an order that makes sense for the purpose at hand.

These types of basic transformations are the essence of generating reports, that is, presenting different views of the same data for different purposes or different audiences. Figure 3 shows the cats in order of their birth dates. Figure 4 shows only the two female cats and presents just their IDs, names, and number of kittens. Each of these reports was generated from the same XML file by applying a different XSL file to its contents.

Resolving Differences Between Disparate XML Data

The DOB data in the Cats XML file is stored in ISO 8601 format, an international standard that defines a method for writing dates and times unambiguously (W3C, 1997). The date part of the standard specifies that a four-digit year be followed by a two-digit month (with a leading zero if necessary) and a two-digit day (again possibly with a leading zero). Hyphens separate the three data items, making `YYYY-MM-DD` the common abbreviation for this format. In addition to being unambiguous, dates expressed in ISO 8601 format have the distinct advantage of being alphabetically sortable. That is, a simple alphanumeric sort will correctly put “2002-01-31” before “2002-02-01,” while the same sort would incorrectly put “January 31, 2002” after “February 1, 2002” and “31.01.2002” after “01.02.2002.”

When a programmer is confronted with the need to compare dates stored in the many formats shown at the beginning of this chapter, one approach is to convert all dates to ISO 8601 format and then do the required comparison (see Figure 5). XSLT provides the capability to do this and similar conversions with relatively few lines of code and surprising execution speed. (Code to do this is presented in the final example of this chapter.) Using XSLT for this task allows a programmer to stay within the XML paradigm, eliminating the need to write functions in other, more general purpose languages.

Communicating Between Applications

The first two sample XSL applications discussed above transform XML into HTML. The third transforms data from one form of XML to another. XSL can actually be used to transform XML data into virtually any text-based format, thus making it an invaluable partner to XML in situations where one application must exchange data with another (Cagle, 2000b).

Consider, for example, the relationships between various components in a multitiered Web application (see Figure 6). A customer might browse product descriptions and prices online using HTML generated from XML. When the customer places the order, that same XML

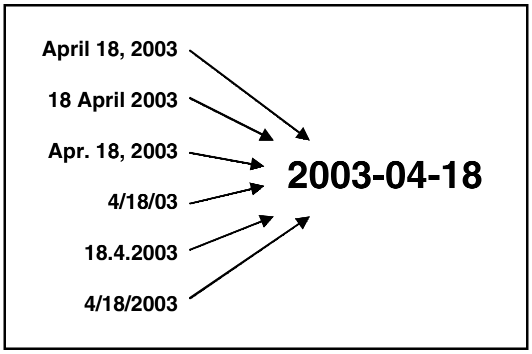


Figure 5: Transforming dates from various disparate forms into a single, standard form: ISO 8601 format.

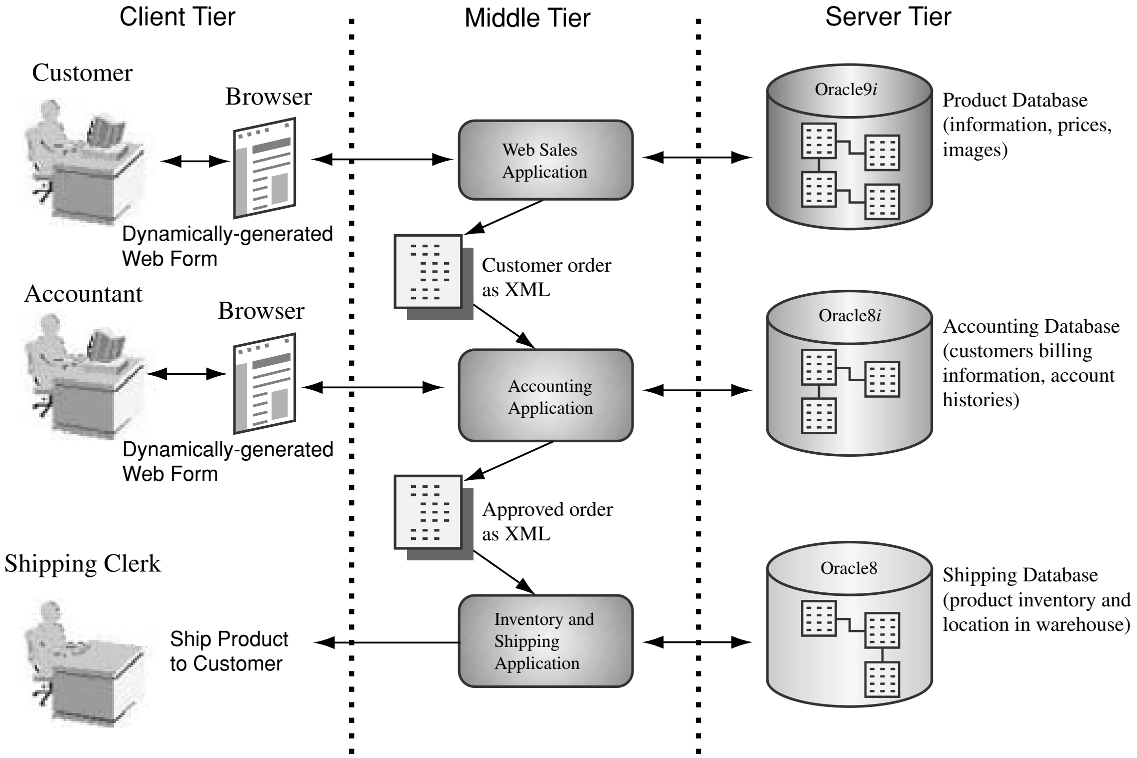

Client Tier
Middle Tier
Server Tier
Customer
Browser
Dynamically-generated Web Form
Accountant
Browser
Dynamically-generated Web Form
Shipping Clerk
Ship Product to Customer
Web Sales Application
Customer order as XML
Accounting Application
Approved order as XML
Inventory and Shipping Application
Oracle9i
Product Database (information, prices, images)
Oracle8i
Accounting Database (customers billing information, account histories)
Oracle8
Shipping Database (product inventory and location in warehouse)

Figure 6: Work and data flow scenario involving XML data in various formats (Oracle, 1999). (Copyright © 1999, 2001, Oracle Corporation. All rights reserved. Used with permission.)

might be processed by another XSL file to generate an SQL query that determines product availability. The accounting department will need the same information in a form compatible with its invoice-generating program, and XSL might provide that in a comma-delimited data stream. The shipping department may require yet another form to generate pick lists and mailing labels. Once again, XSL might be used to perform the required transformation.

WHAT XSL IS AND IS NOT

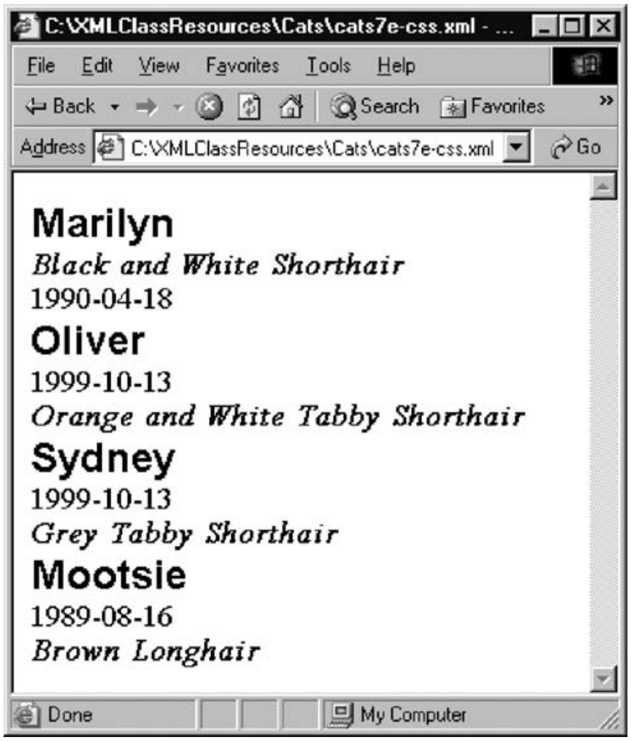
"Decorating" vs. "Transforming" (CSS vs. XSL)

At first glance, XSL may appear to be for XML what CSS is for HTML. Alternatively, a cursory comparison of CSS and XSL might lead one to reasonably—but erroneously—conclude that XSL is a "better" type of CSS. Indeed, one can apply CSS to an XML document to achieve some of the capabilities we have demonstrated with XSL. For example, Figure 7 shows the Cats XML file with CSS attributes applied. However, Table 1 explains the real differences between CSS and XSL. Thus XSL is not a better type of CSS. It is really a different technology with capabilities for manipulating XML data, while CSS is still as valuable as ever for controlling the appearance of generated HTML elements.

XSL as an XML Document

Any discussion of what XSL is and is not will reveal different opinions based on how one uses this technology. However, one issue is not open to debate: XSL structure. An XSL document is first and foremost an XML document, and as such it must conform to all the syntactic and semantic rules for XML documents. This means that if one includes HTML constructs in an XSL file (a common practice), they must be well-formed. All start tags must have end tags or be self-closing (end with `/>` rather than just `>`). Thus one cannot use HTML tags such as `<br>` and `<img>` in an XSL file without their corresponding end tags, `</br>` and `</img>`, which almost never appear in standard HTML. Shortcuts are allowed, such as `<br/>`. (The author has found that `<br/>` [without any spaces] is sometimes interpreted as `<br><br>` by browsers. `<br />` [with a space between the `r` and the `/]` seems to be reliably interpreted as a single `<br>` tag.) One also cannot use entity references such as `&nbsp;`, because XSL has only five predefined entity references: `&lt;`, `&gt;`, `&amp;`, `&apos;`, and `&quot;`. (For `&nbsp;`, you can use `&#160;`.) Finally, if generating HTML, one must be more careful with white space characters, because these are significant in XML while typically insignificant in HTML. These types of issues become moot if one works in XHTML, which is a version of HTML that conforms to XML standards, thus requiring all documents to be well-formed in the XML sense.

The characteristic that clearly distinguishes XSL documents from all other XML documents is their *namespace*. The most recent XSL namespace is http://www.w3.org/1999/XSL/Transform (W3C, 1999), but one may see some older books and papers using a previous version, http://www.w3.org/TR/WD-xsl. Internet Explorer


C:\XMLClassResources\Cats\cats7e-css.xml - ...
File Edit View Favorites Tools Help
Back Search Favorites
Address C:\XMLClassResources\Cats\cats7e-css.xml
Go
Done
My Computer

Marilyn
Black and White Shorthair
1990-04-18
Oliver
1999-10-13
Orange and White Tabby Shorthair
Sydney
1999-10-13
Grey Tabby Shorthair
Mootsie
1989-08-16
Brown Longhair

Figure 7: Cats XML file linked to a CSS file.

5 (including version 5.5) only supports the older version unless one adds a plug-in. Internet Explorer 6.0 supports the newer version, which is more complete and more closely conforms to the W3C Recommendation. The many various versions of all browsers, including Netscape and Mozilla, provide widely varying built-in levels of support for XSL.

The XSL Processing Model

Purists will state that XSL is a *pattern-matching* language, not a *programming* language. Indeed, it would be quite cumbersome to do many general purpose programming tasks in XSL. There are, however, a number of programming constructs built into XSL, such as `if` constructs, `choose` constructs (analogous to `switch` or `select` in other languages), sorting, iteration, recursive descent, and numerous others as well as pattern matching.

The overall strategy, however, is *declarative* rather than procedural or functional. This means that you "specify how you want the result to look rather than saying how it should be transformed" (Anderson et al., 2000, p. 375). An XSL *engine*—software that applies an XSL file to an XML file—first loads an XML source document and an XSL stylesheet into memory (see Figure 8). Internally, each of these documents is represented as a multibranching tree.

The XSL engine then begins processing the stylesheet tree at its root node. The output specifications in this node may cause other stylesheet nodes to be applied to the source tree in turn. Those may be applied to the whole source tree, selected subtrees, or collections of source tree nodes. Each stylesheet node specifies how the transformation result for some part of the source tree is to look. As shown in Figure 8, applying stylesheet specifications to a source tree results in the XSL engine generating a *result tree*. That result tree can then be output in any of a number of formats, of which text, HTML, and XML itself are the most common.

XSL APPLICATION INFRASTRUCTURE

The Minimal XSL Document

As mentioned above, an XSL document is first and foremost an XML document. Thus all XSL document files begin with the standard XML *processing instruction:*

```
<?xml version="1.0" ?>
```

(As of February 16, 2003, there is only one officially accepted version of XML, so the version number is always 1.0. The W3C Recommendation for XML Version 1.1 is currently under review.)

The *root element* of an XSL document is always `stylesheet`. This element name comes from the XSL namespace, so it must be preceded by a *namespace prefix* (W3C, 1999) followed by a colon. By convention, most people use `xsl` as the namespace prefix, declared with an `xmlns` attribute on this root element (see below). In addition, the stylesheet element requires a `version` attribute, which (like XML itself) is currently 1.0. Thus the minimum well-formed and valid XSL document that uses the most recent XSL namespace (as of July 2002) must contain the following three lines:

Table 1 Capabilities of CSS vs. XSL

CSS	XSL
CSS controls the formatting properties of elements, their so-called "decorations"	XSL "transforms" an XML tree into a new tree
CSS cannot reorder elements, generate text, or perform calculations	XSL can do all of these
CSS cannot access XML attributes	XSL has full control over XML attributes, just as it does over elements
CSS can only render a node's value once	XSL can use a node's value as many times as desired and in as many contexts as needed

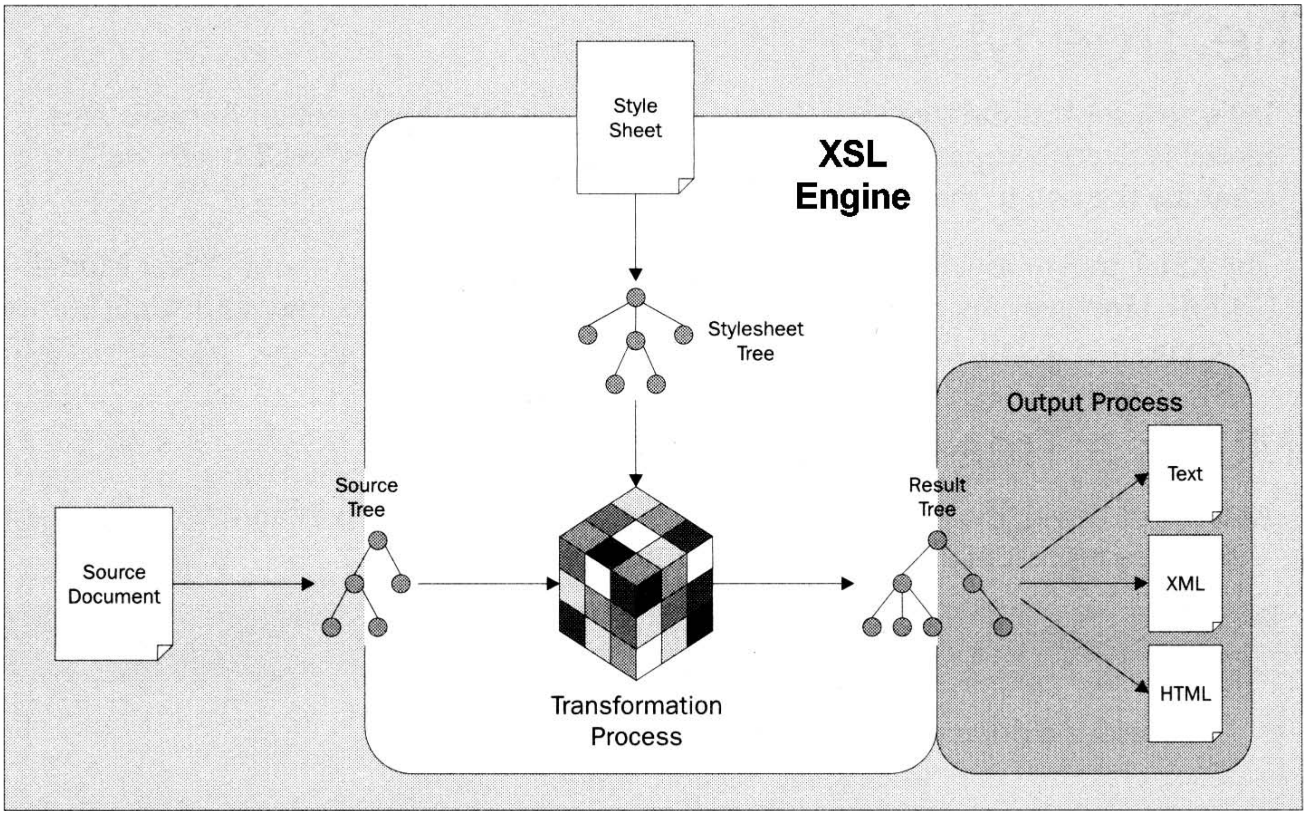


Figure 8: The XSL processing model (Kay, 2000, p. 49). (Copyright © 2000, Wrox Press, Ltd. All rights reserved. Used with permission.)

```
<?xml version="1.0" ?>
<xsl:stylesheet xmlns:xsl=
  "http://www.w3.org/1999/XSL/Transform"
    version="1.0">
    ...
</xsl:stylesheet>
```

XSL Templates

The abstract stylesheet nodes referred to in the discussion of the XSL processing model above are specified using XSL *templates,* which are introduced by the `xsl:template tag`. These can be thought of as pieces of output that get generated—or further specifications that get processed—when the XSL engine state causes those templates to be applied. Each template is differentiated by a *match condition* that identifies the *context* in which it is applied. (Templates with identical match conditions may optionally be differentiated by *modes*.)

The XSL engine starts its processing with the template whose match condition specifies the XSL document's root context. The `template` tag for this node is

```
<xsl:template match="/">
    ...
</xsl:template>
```

It is important to realize that the XSL document root context is *not* the XML document root node. One should think of the XSL root context as an abstract context *just above* the XML root node, as represented in Figure 9. This template is equivalent to the `main` function in a C or C++ or Java program: it is where processing begins. The output specified here is typically code that "frames" output that will be generated by other templates. For an XSL document designed to generate HTML, this means that this "main" template typically contains at least the opening and closing `<html>` and `</html>` tags, most of the output for the `head` section, and the `body` tags. A simple XSL file structured in this way is shown in Listing 1.

If you try to bring up this file in Internet Explorer Version 6.0, you will get the XML tree display shown in Figure 10. This is because an XSL file *is an XML file,* and without *applying* it to an XML file via an XSL *engine,* it is no different than any other XML file.

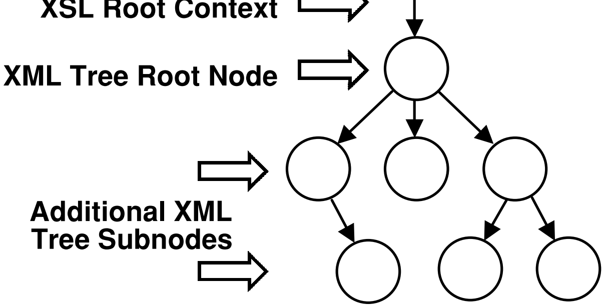
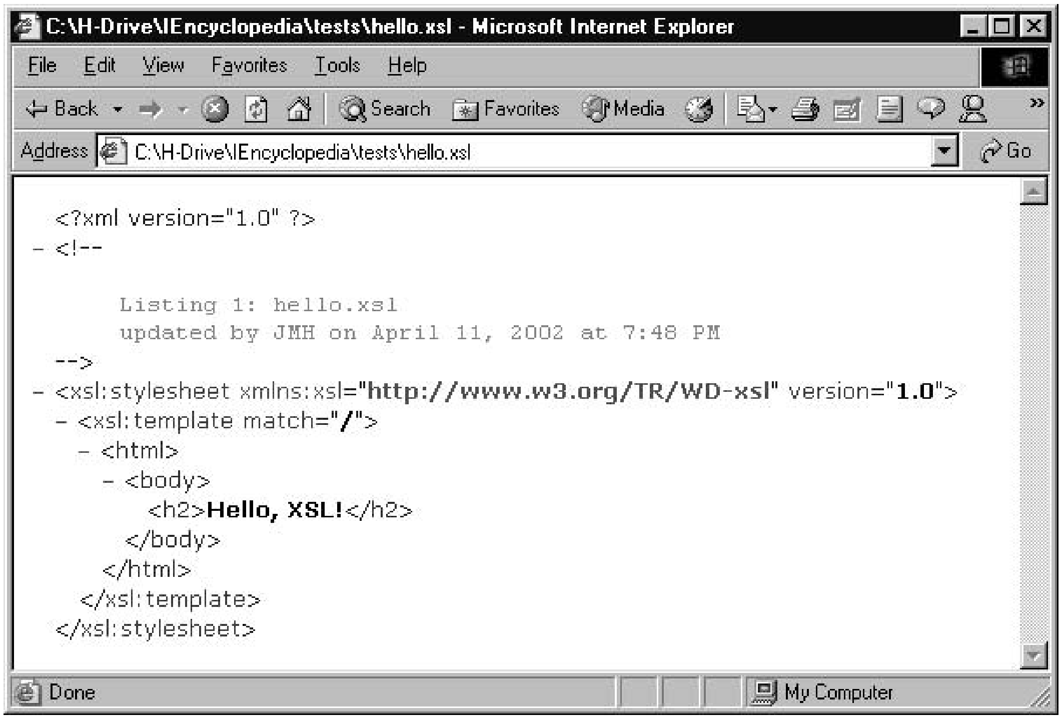


Figure 9: The XSL root context vs. the XML root node.


C:\H-Drive\IEncyclopedia\tests\hello.xsl - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address C:\H-Drive\IEncyclopedia\tests\hello.xsl
Go
Done
My Computer

```
  <?xml version="1.0" ?>
- <!--

      Listing 1: hello.xsl
      updated by JMH on April 11, 2002 at 7:48 PM
  -->
- <xsl:stylesheet xmlns:xsl="http://www.w3.org/TR/WD-xsl" version="1.0">
  - <xsl:template match="/">
    - <html>
      - <body>
          <h2>Hello, XSL!</h2>
        </body>
      </html>
    </xsl:template>
  </xsl:stylesheet>
```

Figure 10: File `hello.xsl` displayed as XML.

Listing 1: File hello.xsl.

```
<?xml version="1.0" ?>
<!--
  hello.xsl - minimal XSL file
  updated by JMH on April 11, 2002
-->
<xsl:stylesheet xmlns:xsl=
"http://www.w3.org/1999/XSL/Transform"
                   version="1.0">

<xsl:template match="/">
   <html>
     <body>
       <h2>
         Hello, XSL!
       </h2>
     </body>
   </html>
 </xsl:template>

 </xsl:stylesheet>
```

Choosing an XSL Engine

When it comes to selecting an XSL engine, there are many choices. We will follow the lead of most other books on this subject by using the XSL engine built into Internet Explorer 6.0 (Microsoft, 2002), because that makes it easy to see our results. As mentioned earlier, the XSL engine in Internet Explorer 6.0 conforms quite well to the latest World Wide Web XSL Recommendation, but the one in previous versions of Internet Explorer does not.

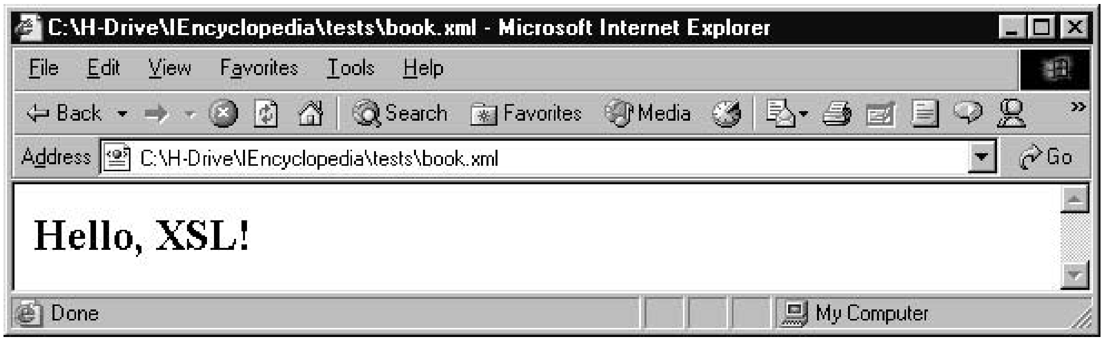
However, one should realize that there are numerous other high quality XSL engines available and several ways to link an XML file to an XSL file using these engines. The choice of which engine and which technique to use depends largely on where one is using XSL (on the client side, the server side, or in a stand-alone application) and the format of one's XML and XSL documents (files on disk or data structures in memory). We cannot explore all the possibilities in this chapter, but it is worth mentioning that the differences in browser versions at the time of this writing make applying XSL on the client side extremely unreliable.

Most of today's XSL processing that generates HTML Web pages is therefore done on the server side. A popular XSL engine that integrates very smoothly with Java Web servers is the Xalan-Java engine, which is available free of charge from the Apache Software Foundation (Apache, 2002). In this chapter we will work with the client-side Internet Explorer engine, but the appendix provides listings of small programs to apply XSL files to XML files on the server side. These programs include a small, self-contained JavaServer Page that demonstrates using hard-coded XML and XSL file names and an analogous Java Servlet that processes XML and XSL file names supplied by an HTML form.

Applying an XSL File to an XML File

To link an XSL file to an XML file, one can include a processing instruction in the XML file that specifies the relative path to the XSL file to be applied:

```
<?xml-stylesheet type="text/xsl"
 href="hello.xsl" ?>
```


C:\H-Drive\IEncyclopedia\tests\book.xml - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address C:\H-Drive\IEncyclopedia\tests\book.xml
Go
Done
My Computer

Hello, XSL!

Figure 11: File `hello.xsl` applied to XML file `hello.xml`.

Since the simple XSL file in Listing 1 makes no reference to the elements in any XML file that may link to it, we can create a minimal XML file that includes only this processing instruction and some arbitrary root tag required to make the XML document well formed. Bringing up the XML file shown in Listing 2 in Internet Explorer Version 6.0 generates the output shown in Figure 11.

Listing 2: Minimal XML file that links to an XSL file.

```
 1 <?xml version="1.0"?>
 2 <!--
 3   hello.xml - minimal version including
 4     xml-stylesheet processing instruction
 5   updated by JMH on April 11, 2002
 6 -->
 7 <?xml-stylesheet type="text/xsl"
 8   href="hello.xsl" ?>
 9 <hello>
10 </hello>
```

Inserting the `xml-stylesheet` processing instruction in an XML file as done at line 7 in Listing 2 is equivalent to hard-coding the link. Microsoft JScript (their extended version of JavaScript) provides the ability to load XML and XSL files dynamically and apply one to the other under program control. This technique allows XML data styled with XSL to be seamlessly integrated into HTML pages. The minimal code to accomplish this is shown in Listing 3. (One could, of course, read the file names into variables and apply them dynamically, but they are shown hard-coded in Listing 3 for simplicity.) Of course, JScript is only one way to apply XSL to be applied to XML dynamically. The appendix includes code to do so through a JavaServer Page and a Java Servlet, and other techniques are possible as well.

EXTRACTING XML DATA

We're now ready to explore the code that gives XSL technology its real power. This code centers around the `xsl:template` element we have already seen, but with considerably more complexity. A large portion of that complexity resides in XPath, the language used to express the crucial *patterns* that appear in the `match` and `test` attributes of XSL elements. Like XSLT, XPath is an integral part of XSL. Of course, entire books have been written about XSLT and XPath (see, for example, the *XSLT Programmer's Reference* by Michael Kay, 2000), but the scope of this chapter is considerably less ambitious. It strives only to give a feel for the types of things one can do with some of the basic XSLT elements, how XPath is used to apply those elements to selected parts of the XML tree, and how one should think of an XSL document as a whole.

We'll use the XML file in Listing 4 for the examples in this section. This file contains information on the titles and authors of four of the chapters in the *Internet Encyclopedia* using a variety of XML structures. These structures allow us to demonstrate how XSL can address each piece of data by applying different XSL files to this XML document. Most of the transformations in the remainder of this chapter were generated under program control using the minimal JScript code shown previously in Listing 3, which explains why there is no `xml-stylesheet` processing instruction hard-coded in the `book.xml` file shown.

Extracting Single Data Items

Extracting Data Stored in Element Nodes

The main way to extract data from an XML document is via the `xsl:value-of` element, and the heart of that element is the `select` attribute which specifies the data to be extracted. The value of the `select` attribute is an XSL *pattern*. The syntax of that pattern is an XPath *expression,* which defines a *collection* of XML nodes for processing by XSLT. Let's replace line 14 in Listing 1 with

```
<xsl:value-of select="book/chapter/
  affiliation/department" />
```

The result is the single string:

```
    Information Technology
```

To understand why, consider the following points:

Listing 3: Minimal JScript code to apply an XSL file to an XML file.

```
<html>
<!--
  Minimum Code for Applying an XSL file to an XML file on the
      Client Side using Internet Explorer Version 6.0
  updated by JMH on April 13, 2002 at 1:59 PM
-->
<body>
  <script type="text/javascript">
    var xmlDoc = new ActiveXObject( "Microsoft.XMLDOM" );
    xmlDoc.async = false;           //disable multithreading
    xmlDoc.load( "hello.xml" ) ;

      var xslStyleSheet = new ActiveXObject( "Microsoft.XMLDOM" );
      xslStyleSheet.async = false;    //disable multithreading
      xslStyleSheet.load( "hello.xsl" ) ;

      document.write( xmlDoc.transformNode( xslStyleSheet ) ) ;
    </script>
 </body>
 </html>
```

Listing 4: XML file for use in examples in this section.

```
<?xml version="1.0"?>
<!--
  book.xml - selected chapters and their authors
  updated by JMH on July 17, 2002 at 02:24 PM
-->
<book>
  <chapter title="Java Server Pages (JSP)">
     <author last="Pratter" first="Frederick" />
     <title>Adjunct Instructor</title>
     <affiliation>University of Montana
       <department>Information Technology</department>
       <e-mail>pratter@cs.umt.edu</e-mail>
     </affiliation>
  </chapter>
  <chapter title="JavaScript">
    <author last="Roussos" first="Constantine" />
    <title>Professor</title>
    <affiliation>Lynchburg College
      <department>Computer Science</department>
      <e-mail>roussos@lynchburg.edu</e-mail>
    </affiliation>
  </chapter>
  <chapter title="Extensible Stylesheet Language (XSL)">
    <author last="Heines" first="Jesse" middle="M."/>
    <title>Associate Professor</title>
    <affiliation>University of Massachusetts Lowell
      <department>Computer Science</department>
      <e-mail>heines@cs.uml.edu</e-mail>
    </affiliation>
    <chapter title="XML, XML Everywhere, But Not a Drop of Compatibility" />
    <chapter title="What XSL Is and Is Not"/>
    <chapter title="XSL Application Infrastructure"/>
    <chapter title="Extracting XML Data"/>
   </chapter>
```

```
35    <chapter title="Extensible Markup Language (XML)">
36      <author last="Ulmer" first="John"/>
37      <title>Assistant Professor</title>
38      <affiliation>Purdue University
39        <department>Computer and Information Systems Technology</department>
40        <e-mail>jjulmer@tech.purdue.edu</e-mail>
41      </affiliation>
42    </chapter>
43  </book>
```

- This `xsl:value` instruction is being executed within the `xsl:template` element whose match attribute is `"/"`. Therefore, as shown previously in Figure 9, the context of this template is the XSL root context, just above the XML root node.
- To "descend into" the XML document, we must therefore first reference the XML root node, `book`.
- To descend further into the document, we refer to the nodes in the order in which they appear in the XML source tree, separating successive tree levels with forward slashes (/). This is basic XPath syntax. The pattern is most easily read from right to left: we are looking for a `department` node that has an `affiliation` node as its parent, a `chapter` node as its grandparent, and a `book` node as its great-grandparent.
- When, as in this case, the XSL pattern references a node whose only child is an unnamed text node (that is, an element whose DTD specification is `<!ELEMENT` *elementName* `(#PCDATA)>`), the `xsl:value-of` element returns the text stored in that node. Thus in this case we get the text "Information Technology."
- Note that in this example the text of only the first node that matches the XPath specification in the `select` attribute is returned. (We will see how to reference groups of nodes a little later.)

Extracting Data Stored in Attribute Nodes

To extract data stored in attributes, we use the @ sign:

```
<p><i>Chapter Title:</i>&#160;
   <xsl:value-of
     select="book/chapter/@title"/>
</p>
```

Reading the pattern from right to left: we are looking for the text stored in an attribute node named `title` that is a child of a `chapter` node which is in turn a child of a `book` node. We've added some additional HTML code to this group of instructions to show further how XSL output can be wrapped in formatting text. The resultant output is:

```
   Chapter Title: Cascading Stylesheets (CSS)
```

Extracting Text Data in Mixed Content Nodes

Look at the structure of the data stored inside the `affiliation` tags in Listing 4. This type of structure is called *mixed content* because it includes both text and subelements. The DTD code for this structure is

```
<!ELEMENT affiliation
  (#PCDATA | department | e-mail)*>
<!ELEMENT department (#PCDATA)>
<!ELEMENT e-mail (#PCDATA)>
```

If we try to extract the text data stored in the `affiliation` node that begins on line 10 in Listing 4 with the statement,

```
<xsl:value-of
  select="book/chapter/affiliation" />
```

the result is all of the text in all of the subelements:

```
   University of Montana Information
     Technology pratter@cs.umt.edu
```

To get only the text at the first level of the `affiliation` node, we use another XPath feature called a *location path* (W3C, 2001). In this case we want to use the `text()` location path, which selects all the text node children of the context node. In the problem at hand, the context node is `affiliation`. So to get just its text, we use the statement:

```
   <xsl:value-of select="book/chapter/
     affiliation/text()" />.
```

This gives us just the text we desire:

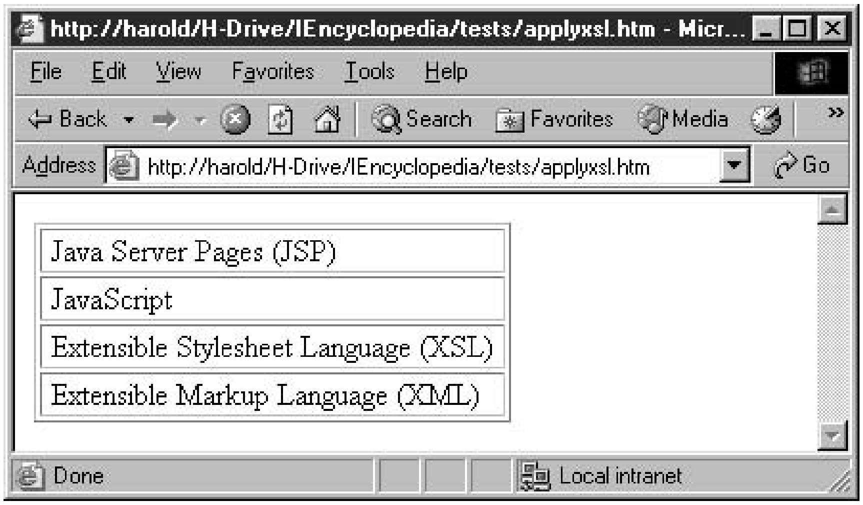
```
    University of Montana
```

We have now seen how to extract data from elements that contain only text, attribute values, and nodes that contain mixed content. These are the three most common situations for any *single* piece of data. Let's now see how to extract sets of data.

Extracting Sets of Data Items

Iteration

One way to extract sets of data is to use the `xsl:for-each` instruction. Like the `xsl:value-of` instruction, `xsl:for-each` has a `select` attribute, but this time the attribute is interpreted as a *node set expression* that selects all the XML data items that match its XPath expression. The `xsl:for-each` instruction then applies the template between its start and end tags to each node in the set.


http://harold/H-Drive/IEncyclopedia/tests/applyxsl.htm - Micr...
File
Edit
View
Favorites
Tools
Help
Back
Search
Favorites
Media
Address
http://harold/H-Drive/IEncyclopedia/tests/applyxsl.htm
Go
Done
Local intranet

Java Server Pages (JSP)
JavaScript
Extensible Stylesheet Language (XSL)
Extensible Markup Language (XML)

Figure 12: Result of applying the XSL iteration construct in Listing 5 to the XML file in Listing 4.

Consider the code in Listing 5. The `xsl:for-each` instruction appears at line 13, and its XPath specification (reading right to left) selects all of the `chapter` elements that are children of `book` elements. This listing also formats the XSL output as an HTML table, a common practice with tabular data. Applying the XSL file in Listing 5 to `book.xml` in Listing 4 results in the display shown in Figure 12.

Listing 5: XSL iteration with the xsl:for-each instruction.

```
 1 <?xml version="1.0" ?>
 2 <!--
 3   book2.xsl
 4   updated by JMH on April 15, 2002
 5 -->
 6 <xsl:stylesheet xmlns:xsl=
 7   "http://www.w3.org/1999/XSL/Transform"
 8  version="1.0">
 9 <xsl:template match="/">
10   <html>
11     <body>
12       <table border="1">
13           <xsl:for-each
14             select="book chapter">
15              <tr>
16              <td>&#160;<xsl:value-of
17                select="@title" />&#160;</td>
18              </tr>
19           </xsl:for-each>
20       </table>
21     </body>
22   </html>
23 </xsl:template>
24 </xsl:stylesheet>
```

This example also demonstrates the important concept of *node context changes*. Note that the XPath expression in the `xsl:value-of` instruction's `select` attribute (line 16 in Listing 5) is `"@title,"` not `"book/chapter/@title"` as in the example we looked at for extracting data stored in attribute nodes:

```
<p><i>Chapter Title:</i>&#160;
   <xsl:value-of
     select="book/chapter/@title"/>
</p>
```

The difference here is that the `xsl:for-each` instruction *changes the context* of the instructions inside its start and end tags to the node specified in its `select` attribute. Thus line 16 is executed in the context of a `book/chapter` node. Saying it another way, line 16 is executed on a `book/chapter` *subtree*. We extract the text in the `title` attribute by referring to the XPath *relative to* the current context. Since we're already at `book/chapter`, we only have to go down one more level to `@title`. Understanding context changes is crucial to understanding the preferred way of extracting sets of data items: using recursive descent.

Recursive Descent

There is nothing wrong with iteration, but it is generally thought of as a procedural construct. Since XSL is declarative by nature, creating additional templates and applying them under certain conditions is more in keeping with XSL's overall design philosophy. Templates can be applied recursively as one descends into the XML source tree. Thus this technique is a form of *recursive descent*.

The XSL instruction used to apply templates is aptly named `xsl:apply-templates`. Like the `xsl:for-each` instruction, `xsl:apply-templates` uses a `select` attribute to specify a set of nodes to which matching templates should be applied.

To change the code in Listing 5 from iteration to recursive descent, we first replace lines 13–19 with the single statement:

```
<xsl:apply-templates
  select="book/chapter" />
```

We then define a new template:

```
<xsl:template match="chapter">
  <tr>
    <td>&#160;<xsl:value-of select=
      "@title" />&#160;</td>
  </tr>
</xsl:template>
```

Note that the value of the match attribute is not only another XSL pattern, but also note that the context of this pattern is somewhat "free floating" and does not include the full XPath expression in the xsl:apply-templates select attribute. In this example, match="chapter" will cause this template to be called *whenever* the context is "chapter," which it is when we specifically descend into the XML document's book node and select all the chapter nodes. The output generated by the xsl:apply-templates and xsl:template instructions is exactly the same as that in Figure 12.

If chapters had subchapters that were also identified with chapter tags, we could do a true recursive descent into the source tree to find all the chapter nodes by changing the xsl:apply-templates instruction's select attribute as follows:

```
<xsl:apply-templates select="//chapter"/>
```

Again reading right to left, this expression tells the XSL engine to select "all chapter nodes that are children of any other node." Since the selection is made recursively, all chapter nodes in the following structure would be processed:

```
<chapter title="Extensible Stylesheet Language (XSL)">
  <author last="Heines" first="Jesse" middle="M." />
  <title>Associate Professor</title>
  <affiliation>University of Massachusetts Lowell
     <department>Computer Science </department>
     <e-mail>heines@cs.uml.edu</e-mail>
  </affiliation>
  <chapter title="XML, XML Everywhere, But Not a Drop of Compatibility" />
  <chapter title="What XSL Is and Is Not" />
  <chapter title="XSL Application Infrastructure" />
  <chapter title="Extracting XML Data" />
</chapter>
```

Filtering

One last variation before we move on: the ability to simulate queries into the XML data by *filtering* the set of extracted data items. To do this, one adds a Boolean expression enclosed within square brackets to an XPath. (Such expressions are more precisely called *predicate expressions* and are an integral part of XPath.) For example

```
<xsl:apply-templates select=
  "book/chapter[not(author/@middle)]"/>
```

returns the set of all chapter nodes that are children of book nodes and whose child author nodes do *not* include a middle attribute. (For our sample XML file, this statement would select the nodes for **Frederick Pratter**, **Constantine Roussos**, and **John Ulmer**.)

It is easy to see that such filters can quickly get very complex. The standard Boolean =, !=, and, or, and not operators exist in XPath, as well as numerous functions such as contains and starts-with for strings. When working with strings, remember that an XSL document is an XML document, so single quotation marks must be included inside double quotation marks, or vice versa, because there is no escape character like "\" in C/C++/Java.

This feature provides some of the capabilities of a WHERE clause in SQL queries. For example,

```
<xsl:apply-templates
      select="book/chapter[starts-with
        (author/@first,'J')]"/>
```

returns the set of all chapter nodes that are children of book nodes and the value of the first name attribute of the child author node begins with the letter J. (For our sample XML file, this statement would select the nodes for **Jesse Heines** and **John Ulmer**.)

Sorting Transformations

Once one knows how to refer to each type of data in an XML source using XPath expressions and iterate over sets of nodes or recurse into the XML tree, one has full access to the XML data and can use XSL to transform it in a myriad of ways. Let's look at sorting as an example. This is accomplished by adding xsl:sort instructions as children of either the xsl:for-each or xsl:apply-templates instructions.

The xsl:sort element has three main attributes:

- select specifies the data on which to sort
- data-type is typically either "text" or "number"
- order is either "ascending" or "descending"

If one includes multiple xsl:sort instructions, the first is taken as the primary sort key, the second as the secondary sort key, etc.

The code in Listing 6 generates a table showing the five chapters sorted primarily on the authors' last names and secondarily on their first names. The output is shown in Figure 13.

Listing 6: Sorting data.

```
 1 <?xml version="1.0" ?>
 2 <!--
 3   book3.xsl
 4   updated by JMH on April 15, 2002 at 11:49 AM
 5 -->
 6 <xsl:stylesheet xmlns:xsl="http://www.w3.org/1999/XSL/Transform" version="1.0">
 7
 8 <xsl:template match="/">
 9   <html>
10     <body>
11       <table border="1">
12         <xsl:apply-templates select="book/chapter">
13            <xsl:sort select="author/@last" data-type="text" order="ascending" />
14            <xsl:sort select="author/@first" data-type="text" order="ascending" />
15         </xsl:apply-templates>
16       </table>
17     </body>
18   </html>
19 </xsl:template>
20
21 <xsl:template match="chapter">
22   <tr>
23      <xsl:apply-templates select="author" />
24      <td>&#160;<xsl:value-of select="@title" />&#160;</td>
25   </tr>
26 </xsl:template>
27
28 <xsl:template match="author">
29    <td nowrap="">
30     &#160;<xsl:value-of select="@last" />,
31     <xsl:value-of select="@first" /> &#32;
32     <xsl:value-of select="@middle" />&#160;
33     </td>
34 </xsl:template>
35
36 </xsl:stylesheet>
```

Note the following points about this code.

Lines 13 and 14: The XPath in the `xsl:sort` instructions' `select` attribute is relative to the context specified in the enclosing `xsl:apply-templates` instruction.

Line 23: We have chosen to create a new template to handle `author` elements. Once again, note the context in the `select` attribute at line 23 and how the context changes in the template at lines 28–34.

Line 29: Remember once again that an XSL document is an XML document, so we cannot use the standard HTML `<td nowrap>` construct because every attribute must have a value. Thus we have set the value to something. An empty string will do just fine: `<td nowrap="">`. (In XHTML, the `td` element has no `nowrap` attribute. In that case, you must use CSS to achieve the same result: `<td style="white-space: nowrap">`.)

Lines 30–32: If we put nothing at the end of line 31, the authors' first and middle names will be run together. In line 24 we used `&#160;`, the XSL equivalent of `&nbsp;`, to get the extra aesthetically pleasing spaces before and after the text in each cell of our table. But if we use `&#160;` at the end of line 31, we get two spaces instead of one. Thus we have used `&#32;`, which is the standard ASCII space character.

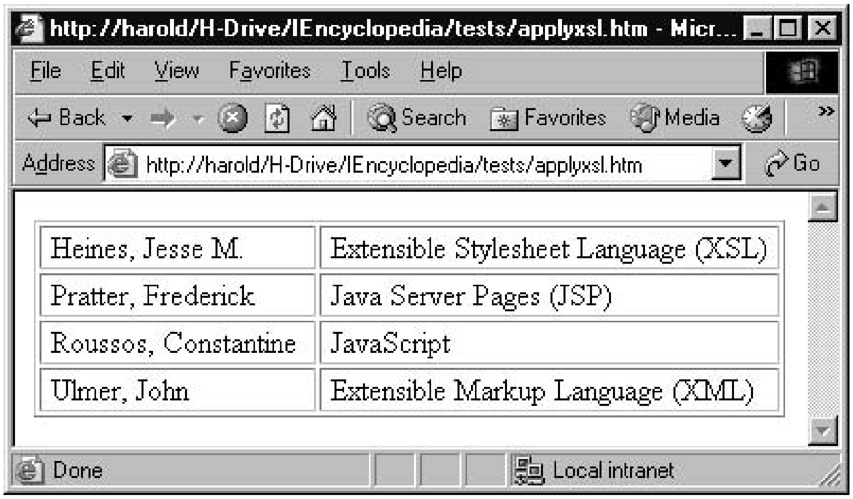
CONTROLLING XSL PROCESSING ENGINE FLOW

Calling Templates with Parameters

The `xsl:for-each` and `xsl:apply-templates` instructions certainly provide some degree of flow control within the XSL processing engine. Further control can be achieved with the `xsl:call-template` instruction, which allows XSL templates to be called by name just like standard subroutines. In a nutshell, the syntax is

```
<xsl:call-template name="templateName" />
```

Rather than the `match` attribute we saw in templates called by `xsl:apply-templates`, templates called by name have a `name` attribute:


http://harold/H-Drive/IEncyclopedia/tests/applyxsl.htm - Micr...
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://harold/H-Drive/IEncyclopedia/tests/applyxsl.htm
Go
Done
Local intranet

Heines, Jesse M.	Extensible Stylesheet Language (XSL)
Pratter, Frederick	Java Server Pages (JSP)
Roussos, Constantine	JavaScript
Ulmer, John	Extensible Markup Language (XML)

Figure 13: Result of applying the XSL sort construct in Listing 6 to the XML file in Listing 4.

```
<xsl:template name="templateName">
  instructions to execute when this template is called
</xsl:template>
```

Another important difference between *applying* and *calling* templates is that in the former the node context changes, as we just saw, while in the latter it does not.

There is also a *parameter* construct that allows templates to be called with values that can be further used to control program flow. This construct can be used with `xsl:apply-templates`, but it is more commonly found with `xsl:call-template`:

```
<xsl:call-template name=" templateName">
  <xsl:with-param name=" parameterName"
    select=" pattern" />
</xsl:call-template>

<xsl:template name=" templateName">
  <xsl:param name=" parameterName" />
    <!-- name must match that used above -->
      instructions to execute when this template is
        called, parameter can be referenced using
        $parameterName
</xsl:template>
```

Examples of these constructs appear in the code for the next major section.

One can therefore see that even though no one would claim that XSL is a general purpose programming language, it is a rich set of instructions that provides many of the basic features of a declarative programming language, coupled with exceptional abilities to manipulate XML data.

Conditional Execution

One of the basic flow control features in any language is the Boolean `if` construct that provides conditional instruction execution. XSL does indeed have an `xsl:if` instruction. Its basic format is

```
<xsl:if test="Boolean expression">
   instructions to execute if the test of the enclosing
     xsl:if is true
 </xsl:if>.
```

This instruction is not as heavily used as one might expect, however, because it has no “else” clause. People therefore tend to use the XSL equivalent of the C/C++/Java `switch` statement: the `xsl:choose` instruction. The basic format of this instruction is

```
<xsl:choose>
  <xsl:when test="Boolean expression">
    instructions to execute if the test of the enclosing
    xsl:when is true
  </xsl:when>
  ... any number of additional xsl:when elements
    may be included here...
  <xsl:otherwise>
    instructions to execute if no other Boolean
      expressions in this xsl:choose are true
  </xsl:otherwise>
</xsl:choose>
```

As you surely suspect by now, the Boolean expressions include XPath expressions, which can, in turn, include Boolean operators and numerous functions. The following sample `xsl:choose` construct outputs **Dear Ms. (*last name*):** if the current context node’s first attribute contains the string “Elizabeth;” otherwise it outputs **Dear Mr. (*last name*):**. Note the use of single and double quotation marks in the `test` attribute of the first `xsl:when` element.

```
<xsl:choose>
  <xsl:when test="@first='Elizabeth'">
     Dear Ms. <xsl:value-of select="@last"
       />:
  </xsl:when>
  <xsl:otherwise>
     Dear Mr. <xsl:value-of select="@last"
       />:
     </xsl:otherwise>
</xsl:choose>
```

One note of clarification to C/C++/Java programmers: Each case (xsl:when or xsl:otherwise) is mutually exclusive. That is, each case ends with an implicit break statement (in C/C++/Java terms), and processing exits the xsl:choose structure as soon as any case is completed. It is therefore critical that tests be sequenced from the most specific to the most general, with xsl:otherwise (if present) as the last case in the sequence.

MAKING DATA IN XML DOCUMENTS COMPATIBLE

To put everything together that we've seen so far, we return to the problem introduced at the beginning of this chapter: incompatibility of date formats. The XML file in Listing 7 has six person elements, each with a birth date specified in one of three different formats: in attributes (line 6), in subelements (lines 10–14), or as text (line 17). (Having different date formats in the same XML file is certainly contrived for this demonstration, but it is common to have different formats in different files, as discussed at the beginning of this chapter.) The XSL file in Listing 8 determines how the birth date for each person is stored and executes the required instructions to transform each into ISO 8601 format. The table resulting from the transformation is shown in Figure 14. Comments on specific XSL techniques follow the listings.

Listing 7: XML file containing birth dates in various formats.

```
 1 <?xml version="1.0"?>
 2 <!-- birthdates.xml
 3   updated by JMH on April 15, 2002 -->
 4 <family>
 5   <person id="Dad">
 6     <birthdate month="November" day="28"
 7       year="1916" />
 8   </person>
 9   <person id="Mom">
10     <birthdate>
11       <month>December</month>
12       <day>13</day>
13       <year>1918</year>
14     </birthdate>
15   </person>
16   <person id="Judy">
17     <birthdate>1/14/42</birthdate>
18   </person>
19   <person id="Carol">
20     <birthdate>
21       <month>July</month>
22       <day>9</day>
23       <year>1943</year>
24     </birthdate>
25   </person>
26   <person id="Henry">
27     <birthdate month="July" day="18"
28       year="1945" />
29   </person>
30   <person id="Jesse">
31     <birthdate>4/18/48</birthdate>
32   </person>
33  </family>
```

Listing 8: XSL file to display birth dates supplied in various formats in ISO 8601 format.

```
34 <?xml version="1.0" ?>
35 <!--
36   birthdates1.xsl
37   updated by JMH on April 15, 2002 at 7:40 PM
38 -->
39 <xsl:stylesheet xmlns:xsl="http://www.w3.org/1999/XSL/Transform" version="1.0">
40
41 <xsl:template match="/">
42   <html>
43     <body>
44       <table border="1">
45         <xsl:apply-templates select="family/person" />
46       </table>
47     </body>
48   </html>
49 </xsl:template>
50
51 <xsl:template match="person">
52   <tr>
53     <td>&#160;<xsl:value-of select="@id"/>&#160;</td>
54     <td><xsl:apply-templates select="birthdate" /></td>
55   </tr>
56 </xsl:template>
```

```

<xsl:template match="birthdate">
  <xsl:choose>

    <!-- handle case where date is in attributes -->
    <xsl:when test="@month">
      <xsl:value-of select="@year"/>-<null/>
      <xsl:call-template name="month">
        <xsl:with-param name="monthName" select="@month" />
      </xsl:call-template>-<null/>
      <xsl:if test="@day &lt; 10">0</xsl:if>
      <xsl:value-of select="@day"/>
    </xsl:when>

    <!-- handle case where date is in subelements -->
    <xsl:when test="month">
      <xsl:value-of select="year" />-<null/>
      <xsl:call-template name="month">
        <xsl:with-param name="monthName" select="month" />
      </xsl:call-template>-<null/>
      <xsl:if test="day &lt; 10">0</xsl:if>
      <xsl:value-of select="day" />
    </xsl:when>

    <!-- handle case where date is in text -->
    <xsl:when test="contains( ./text(), '/' )">
      <!-- extract month, day, and year values as strings -->
        <xsl:variable name="strMonth"
              select="substring-before( ./text(), '/' )" />
        <xsl:variable name="strDay"
              select="substring-before( substring-after(./text(), '/' ), '/' )" />
        <xsl:variable name="strYear"
              select="substring-after( substring-after( ./text(), '/' ), '/' )" />
      <!-- output year, prefixing with "19" if it's expressed in two digits -->
        <xsl:if test="string-length( $strYear ) = 2">19</xsl:if>
        <xsl:value-of select="$strYear" />-<null />
      <!-- output month, prefixing with "0" if it's expressed in one digit -->
        <xsl:if test="$strMonth &lt; 10">0</xsl:if>
        <xsl:value-of select="number( $strMonth )"/>-<null/>
      <!-- output day, prefixing with "0" if it's expressed in one digit -->
        <xsl:if test="$strDay &lt; 10">0</xsl:if>
        <xsl:value-of select="number($strDay)" />
    </xsl:when>

    <!-- handle default case -->
    <xsl:otherwise>
      unknown
    </xsl:otherwise>

  </xsl:choose>
</xsl:template>

 <xsl:template name="month">
   <xsl:param name="monthName"/>
   <xsl:choose>
     <xsl:when test="$monthName='January'">01</xsl:when>
     <xsl:when test="$monthName='February'">02</xsl:when>
     <xsl:when test="$monthName='March'">03</xsl:when>
     <xsl:when test="$monthName='April'">04</xsl:when>
     <xsl:when test="$monthName='May'">05</xsl:when>
     <xsl:when test="$monthName='June'">06</xsl:when>
```

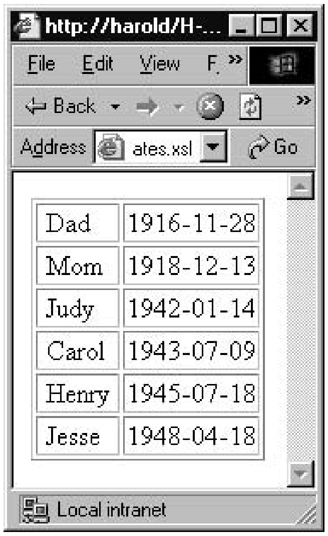
```
118       <xsl:when test="$monthName='July'">07</xsl:when>
119       <xsl:when test="$monthName='August'">08</xsl:when>
120       <xsl:when test="$monthName='September'">09</xsl:when>
121       <xsl:when test="$monthName='October'">10</xsl:when>
122       <xsl:when test="$monthName='November'">11</xsl:when>
123       <xsl:when test="$monthName='December'">12</xsl:when>
124     </xsl:choose>
125   </xsl:template>
126
127   </xsl:stylesheet>
```

The heart of this transformation is in the template that begins on line 58. This template is executed in the context of a `birthdate` node, which exists for each `person` regardless of the birth date's format.

At line 62 we test for the existence of a `month` attribute associated with the `birthdate` node. If such an attribute exists, we assume that the birth date is stored in attributes and proceed accordingly. We extract the value of the `year` attribute and add it to the generated output at line 63. ISO 8601 format specifies that the year value be followed by a hyphen (`YYYY-MM-DD`). However, adding that hyphen at the end of line 63 generates an extra space in the output, so we follow it with a tag that we know HTML processors will ignore: `<null/>`. This dummy tag is immediately followed by another tag at line 64, so no additional spaces are generated.

At line 64 we call the template named `month`, passing it parameter `monthName` with the value extracted from the `month` attribute (line 65). The template named `month` that begins at line 109 transforms month name strings into numbers to conform to ISO 8601 format. Line 67 tests the value of the `day` attribute to see if it is numerically less than 10. If so, this line outputs a 0 to add a leading 0 to the day number to conform to ISO 8601 format.

Line 72 tests for the existence of a `month` subelement in the `birthdate` context. If it exists, we assume that the birth date is stored in subelements and proceed accordingly. The code here is very similar to that in the previous case, except that we extract data from elements rather than from attributes, so no @ signs appear in the `select` and `test` attributes of the various instructions.



Dad	1916-11-28
Mom	1918-12-13
Judy	1942-01-14
Carol	1943-07-09
Henry	1945-07-18
Jesse	1948-04-18

Figure 14: Birthdates transformed to ISO 8601 format.

Line 82 tests whether the text in a `birthdate` element contains a forward slash (`"/"`). If it does, we assume that the birth date is stored in the common American month/day/year format. We then use the XPath string functions `substring-before` and `substring-after` to isolate each of the numbers delineated by the slashes and transform them appropriately to conform to ISO 8601 format.

The template that begins at line 109 is essentially one large `xsl:choose` instruction. Line 110 accepts the `monthName` parameter passed to this template via the `xsl:with-param` instructions at lines 65 and 75. We then use that parameter in the `xsl:when` test attributes by preceding its name with a dollar sign ($). We check for each of the 12 month names in turn and add the corresponding month number (with a leading 0 if necessary) to the output when one of the Boolean tests is true.

Thus the disparate date formats are all made compatible with one another. Of course, one could add additional cases to the `birthdate` template that begins at line 58 to handle all of the variations presented at the beginning of this chapter. (Serial dates require computing a unique day number, where January 1, 1900, is defined as day 1.)

The previous example is fine for generating HTML output as we have been doing throughout this chapter, but many applications that use XML and XSL have no interest at all in generating HTML. Instead, they want to take an XML file in one format and convert it to an equivalent XML file in another format. That is, they would want to convert the file in Listing 7 to a new file with all dates in the same format. The XSL file in Listing 9 does precisely this, generating the output in Listing 10.

XSL FORMATTING OBJECTS

The third component of XSL is the set of *formatting objects* (XSL-FO) that can be used to render output in sophisticated formats such as PostScript, Portable Document Format (PDF), and even Java (using the Abstract Windowing Toolkit for screen display). Use of these formatting objects is a direct extension of the concepts and techniques presented in this chapter, applied using the `http://www.w3.org/1999/XSL/Format` namespace. This namespace provides tags similar to, but considerably more sophisticated than, those found in CSS. To include the XSL-FO namespace in an XSL file, simply add it to the `xsl:stylesheet` tag:

```
<xsl:stylesheet version="1.0"
      xmlns:xsl=
       "http://www.w3.org/1999/XSL/Transform"
      xmlns:fo=
       "http://www.w3.org/1999/XSL/Format">
```

The full set of XSL-FO tags and allowable attributes is truly immense. One "short reference" that merely lists them all is 26 pages long! Thus this chapter can only address the basic structure of an XSL file that uses the XSL-FO namespace. For detailed discussion of XSL-FO, please see the *W3C Recommendation for XSL Version 1.0* (W3C, 2001).

Listing 9: XSL file to generate a new XML file with birth dates in various formats converted to ISO 8601 format.

```
<?xml version="1.0" ?>
<!--
 birthdates2.xsl
 updated by JMH on July 17, 2002 at 03:02 PM
-->
<xsl:stylesheet xmlns:xsl="http://www.w3.org/1999/XSL/Transform" version="1.0">

<xsl:output method="xml" version="1.0" indent="yes" />

 <xsl:template match="/" >
  <xsl:comment> <!-- text within xsl:comment tags will appear in the output file -->
   birthdates2.xml
   generated from birthdates.xml by birthdates2.xsl
  </xsl:comment>
  <xsl:element name="family"> <!-- root element -->
   <xsl:apply-templates select="family/person"/>
  </xsl:element>
 </xsl:template>

 <xsl:template match="person">
  <xsl:text> <!-- go to a new line and add indent two spaces -->
  </xsl:text>
  <xsl:element name="person"> <!-- person subelement -->
   <xsl:attribute name="id"> <!-- first attribute on the person element -->
    <xsl:value-of select="@id"/>
   </xsl:attribute>
   <xsl:attribute name="birthdate">  <!-- second attribute on the person element -->
    <xsl:apply-templates select="birthdate" />
   </xsl:attribute>
  </xsl:element>
 </xsl:template>

 <xsl:template match="birthdate">
  <xsl:choose>

   <!-- handle case where date is in attributes -->
   <xsl:when test="@month">
    <xsl:value-of select="@year"/>
    <xsl:text>-</xsl:text> <!-- use this element to avoid whitespace problems -->
    <xsl:call-template name="month">
     <xsl:with-param name="monthName" select="@month" />
    </xsl:call-template>
    <xsl:text>-</xsl:text>
    <xsl:if test="@day &lt; 10">
     <xsl:text>0</xsl:text>
    </xsl:if>
    <xsl:value-of select="@day" />
   </xsl:when>

  <!-- handle case where date is in subelements -->
  <xsl:when test="month">
    <xsl:value-of select="year" />
    <xsl:text>-</xsl:text>
    <xsl:call-template name="month">
      <xsl:with-param name="monthName" select="month"/>
    </xsl:call-template>
    <xsl:text>-</xsl:text>
    <xsl:if test="day &lt; 10">
      <xsl:text>0</xsl:text>
    </xsl:if>
    <xsl:value-of select="day"/>
    </xsl:when>

    <!-- handle case where date is in text -->
    <xsl:when test="contains( ./text(), '/' )">
      <!-- extract month, day, and year values as strings -->
        <xsl:variable name="strMonth"
             select="substring-before( ./text(), '/' )" />
        <xsl:variable name="strDay"
             select="substring-before( substring-after( ./text(), '/' ), '/' )" />
        <xsl:variable name="strYear"
             select="substring-after( substring-after( ./text(), '/' ), '/' )" />
      <!-- output year, prefixing with "19" if it's expressed in two digits -->
        <xsl:if test="string-length( $strYear ) = 2">19</xsl:if>
        <xsl:value-of select="$strYear"/>
        <xsl:text>-</xsl:text>
      <!-- output month, prefixing with "0" if it's expressed in one digit -->
        <xsl:if test="$strMonth &lt; 10">0</xsl:if>
        <xsl:value-of select="number( $strMonth )" />
        <xsl:text>-</xsl:text>
      <!-- output day, prefixing with "0" if it's expressed in one digit -->
        <xsl:if test="$strDay &lt; 10">0</xsl:if>
        <xsl:value-of select="number($strDay)" />
  </xsl:when>

  <!-- handle default case -->
  <xsl:otherwise>
    unknown
  </xsl:otherwise>

  </xsl:choose>
</xsl:template>

<xsl:template name="month">
  <xsl:param name="monthName" />
  <xsl:choose>
   <xsl:when test="$monthName='January'">01</xsl:when>
   <xsl:when test="$monthName='February'">02</xsl:when>
   <xsl:when test="$monthName='March'">03</xsl:when>
   <xsl:when test="$monthName='April'">04</xsl:when>
   <xsl:when test="$monthName='May'">05</xsl:when>
   <xsl:when test="$monthName='June'">06</xsl:when>
   <xsl:when test="$monthName='July'">07</xsl:when>
   <xsl:when test="$monthName='August'">08</xsl:when>
   <xsl:when test="$monthName='September'">09</xsl:when>
   <xsl:when test="$monthName='October'">10</xsl:when>
   <xsl:when test="$monthName='November'">11</xsl:when>
   <xsl:when test="$monthName='December'">12</xsl:when>
  </xsl:choose>
</xsl:template>
```

```

</xsl:stylesheet>
```

Listing 10: XML file generated by the XSL file in Listing 9.

```
<?xml version="1.0" encoding="utf-8"?>
<!--
   birthdates2.xml
   generated from birthdates.xml by
     birthdates2.xsl
-->
<family>
  <person id="Dad" birthdate=
    "1916-11-28"/>
  <person id="Mom" birthdate=
    "1918-12-13"/>
  <person id="Judy" birthdate=
    "1942-01-14"/>
  <person id="Carol" birthdate=
    "1943-07-09"/>
  <person id="Henry" birthdate=
    "1945-07-18"/>
  <person id="Jesse" birthdate=
    "1948-04-18"/>
</family>
```

XSL formatting objects specify page layout using a hierarchical system of *pages*, *flows*, and *blocks*. Further refinements of the hierarchy include *rectangles*, *borders*, and *spaces*. Full discussion of these topics is beyond the scope of this chapter, but basically one may think of a simple printed document as having a *master layout* that is broken down into *page sequences* that contain *flows* that are composed of *blocks*.

Formatting Static Text

As a first example, consider the code in Listing 11, which is adapted from one of the examples provided with the transformation engine used to render the examples in this section: the Formatting Objects Processor (FOP) from the Apache Group. (This software can be downloaded free of charge from http://xml.apache.org/fop/.) The adapted example renders a version of the abstract for this chapter in Portable Document Format, which can be viewed in the popular Adobe Acrobat Reader as shown in Figure 15. (The explanation of this code that follows is drawn largely from the comments embedded in the original example.)

Listing 11: Simple XSL-FO to render the abstract for this chapter.

```
<?xml version="1.0" encoding="utf-8"?>

<!--
  j2a.fo, adapted from fop-0.20.3\docs\examples\fo\simple.fo
  primary source: http://xml.apache.org/fop/index.html
  updated by JMH on July 18, 2002 at 01:18 PM
-->

<fo:root xmlns:fo="http://www.w3.org/1999/XSL/Format">

  <!-- master document layout -->
  <fo:layout-master-set>
    <fo:simple-page-master master-name="simple"
         page-height="11.0in"  margin-top="1in"     margin-left="1.25in"
         page-width="8.5in"    margin-bottom="1in"  margin-right="1.25in">
      <fo:region-body margin-top="0in"/>
      <fo:region-before extent="0.25in"/>
      <fo:region-after extent="0.5in"/>
    </fo:simple-page-master>
  </fo:layout-master-set>

  <!-- beginning of a page within a document -->
  <fo:page-sequence master-reference="simple">

    <!-- beginning of a flow within a page -->
    <fo:flow flow-name="xsl-region-body">

      <!-- title -->
      <fo:block
              font-size="18pt"      line-height="24pt"       background-color="black"
              font-family="Times"   space-after.optimum="24pt"   color="white"
```

```
32               text-align="center"   padding-top="0pt">
33           The Extensible Stylesheet Language (XSL)
34         </fo:block>
35
36         <!-- subtitle -->
37         <fo:block font-size="14pt"     line-height="20pt"
38                   font-family="Times"  space-after.optimum="12pt"
39                   font-weight="bold"   padding-top="0pt">
40            Abstract
41       </fo:block>
42
43       <!-- first paragraph -->
44       <fo:block font-size="12pt"     line-height="15pt"         text-indent="25pt"
45                 font-family="Times"  space-after.optimum="12pt" text-align="justify">
46         XSL - the Extensible Stylesheet Language - is an XML-based technology for
47         transforming XML documents from one form to another. It uses a declarative
48         programming paradigm and a specific XML namespace that gives programmers full
49         access to all XML components - elements, attributes, and text - and the ability
50         to manipulate them in ways that go far beyond the capabilities of Cascading
51         Style Sheets (CSS).  XSL can be used to control the rendition of XML data,
52         selectively filter the data items selected for transformation, convert data from
53         various incompatible forms into a single, standard form, or implement just about
54         any other operation that one might want to perform on XML data without changing
55         the original XML source. Various parts of XSL are now industry standards (known
56         as World Wide Web Consortium "Recommendations"), and are therefore highly usable
57         even in today's ever-changing Web environment.
58       </fo:block>
59
60       <!-- second paragraph -->
61       <fo:block font-size="12pt"     line-height="15pt"         text-indent="25pt"
62                 font-family="Times"  space-after.optimum="12pt" text-align="justify">
63         This chapter presents the basic concepts and techniques used in XSL. It
64         provides a variety of examples of XSL Transformations (XSLT), XPath Expressions
65         (the language used to refer to collections of XML nodes for processing by XSLT),
66         and XSL Formatting Objects (XSL FO).
67       </fo:block>
68
69      </fo:flow>
70    </fo:page-sequence>
71 </fo:root>
```

We see from line 1 that this is an XML file, but it does not use the XSL transformation engine. Therefore, only the XSL-FO namespace is specified on the fo:root element in line 9. (The next example will use both namespaces as just discussed.)

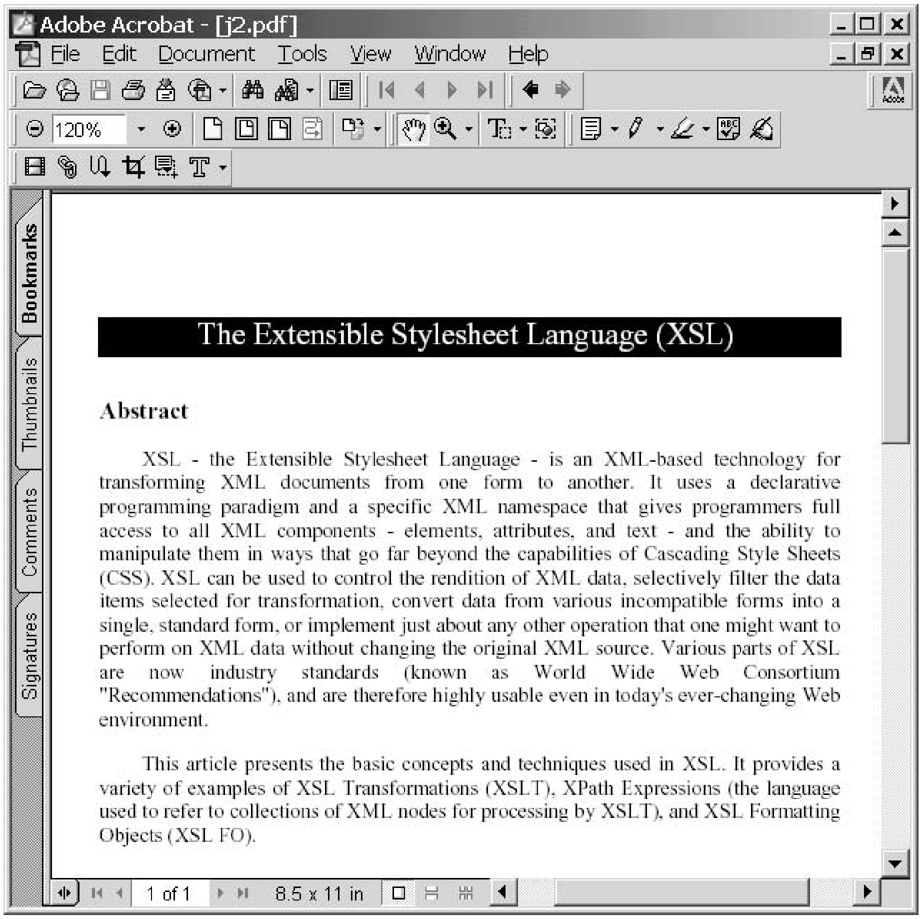
The fo:root element must contain one and only one fo:layout-master-set element (line 12), which in turn contains one or more page master elements that specify sets of *master layout* parameters. In this example, there is only one fo:simple-page-master element that defines the layout for all pages. If there were multiple page layouts, they would be differentiated by their master-name attributes (line 13).

A document's content is organized into one or more *page sequences*. Each fo:page-sequence element (line 23) has a master-reference attribute whose value corresponds to the master-name attribute of one of the previously defined page master elements (line 13).

Page sequences contain *flows*, which can be positioned in one of five regions: the page header or footer, the left or right margin, or the body. The fo:flow tag at line 26 begins the definition of a body flow.

Actual content is then contained within *blocks* that contain text and formatting instructions. There are four blocks in this example. The first block, which begins at line 29, displays the overall title centered in 18-point white type on a black background (72 points = 1 inch). The text for this block appears on line 33. The second block, which begins at line 37, displays a subtitle left-justified in bold 14-point type. The text for this block appears on line 40.

The format of the two paragraphs is defined at lines 44–45 and 61–62. The first line of each paragraph is indented 25 points and their text is justified (all text in all lines except the first is aligned at both the left and right margins).



The Extensible Stylesheet Language (XSL)

Abstract

XSL - the Extensible Stylesheet Language - is an XML-based technology for transforming XML documents from one form to another. It uses a declarative programming paradigm and a specific XML namespace that gives programmers full access to all XML components - elements, attributes, and text - and the ability to manipulate them in ways that go far beyond the capabilities of Cascading Style Sheets (CSS). XSL can be used to control the rendition of XML data, selectively filter the data items selected for transformation, convert data from various incompatible forms into a single, standard form, or implement just about any other operation that one might want to perform on XML data without changing the original XML source. Various parts of XSL are now industry standards (known as World Wide Web Consortium "Recommendations"), and are therefore highly usable even in today's ever-changing Web environment.

This article presents the basic concepts and techniques used in XSL. It provides a variety of examples of XSL Transformations (XSLT), XPath Expressions (the language used to refer to collections of XML nodes for processing by XSLT), and XSL Formatting Objects (XSL FO).

Figure 15: PDF rendering of the XSL-FO file in Listing 11.

Formatting Text from an XML Document

The real power of XSL formatting objects, of course, comes into play when they are combined with XSL transformations to format text read dynamically from XML files. The basic structure of the XSL file is the same as that shown for static text in the first example, but the XSL-FO tags are now embedded within XSLT tags that control flow and pull data from an XML file. That may sound more like alphabet soup than computer programming, but the example that follows is intended to clarify how these technologies fit together.

This example uses the XML file in Listing 12, which is an excerpt from the glossary for this chapter. The corresponding XSL file is in Listing 13.

First, we see the presence of both namespaces defined in the `xsl:stylesheet` element at lines 81–83. Lines 85–87 are the standard top level "main" template, which simply passes processing on to the XML document's root element, `article`.

The template that matches XML element `article` begins at line 89. It contains the `fo:root` element (line 90) and defines the *master layout* using an `fo:layout-`

Listing 12: Excerpt from the glossary for this chapter encoded in XML.

```
 1 <?xml version="1.0" ?>
 2 <!--
 3   jglossary.xml, adapted from fop-0.20.3\docs\examples\markers\glossary.xml
 4   primary source: http://xml.apache.org/fop/index.html
 5   Jesse M. Heines, UMass Lowell Computer Science, heines@cs.uml.edu
 6   updated by JMH on July 17, 2002 at 10:04 PM
 7 -->
 8 <article>
 9   <title>The Extensible Stylesheet Language (XSL)</title>
10   <section>
11     <title>Glossary</title>
```

```
   <entry>
     <term>context</term>
      <definition>
        as used in XSL, the tree level at which an instruction is executed; a single
        instruction will produce different results if executed in different contexts
      </definition>
    </entry>
    <entry>
      <term>declarative language</term>
      <definition>
        a computer language in which one specifies desired results rather than the
        procedures used to achieve those results (compare to pattern matching
        language and procedural language)
      </definition>
    </entry>
    <entry>
      <term>engine</term>
      <definition>
        as used in XSL, a program that applies the declarations in an XSL file to the
        data in an XML file by performs the actions necessary to achieve the
        specified transformation
      </definition>
    </entry>
    <entry>
      <term>entity reference</term>
      <definition>
        a symbol in an XML file that begins with an ampersand and ends with a semi-
        colon; there are only five built-in entities in XSL, while there are many
        more in XML
      </definition>
    </entry>
    <entry>
      <term>eXtensible Stylesheet Language (XSL)</term>
      <definition>
        an XML-structured technology that uses XSL transformations (XSLT), XPath
        addressing schemes, and (optionally) XSL formatting objects (XSL FO) to
        transform XML data into other forms
      </definition>
    </entry>
    <entry>
      <term>filtering</term>
      <definition>
        the process of extracting selected data from an XML file that meet a set of
        specified criteria
      </definition>
    </entry>
    <entry>
      <term>match condition</term>
      <definition>
        a template attribute that specifies the engine state in which that
        template's instructions will be executed
      </definition>
    </entry>
    <entry>
      <term>mode</term>
      <definition>
        a further refinement of a match condition that allows differentiation
        between multiple templates with the same match condition
      </definition>
    </entry>
```

```
  </section>
</article>
```

Listing 13: XSL-FO code to render the glossary XML file.

```
<?xml version="1.0" encoding="utf-8"?>
<!--
  jglossary.xsl, adapted from fop-0.20.3\docs\examples\markers\glossary.xsl
  primary source: http://xml.apache.org/fop/index.html
  Jesse M. Heines, UMass Lowell Computer Science, heines@cs.uml.edu
  updated by JMH on July 18, 2002 at 08:55 AM
-->
<xsl:stylesheet version="1.0"
  xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
  xmlns:fo="http://www.w3.org/1999/XSL/Format">

<xsl:template match="/">
  <xsl:apply-templates select="article"/>
</xsl:template>

<xsl:template match="article">
  <fo:root xmlns:fo="http://www.w3.org/1999/XSL/Format">

   <fo:layout-master-set>
     <fo:simple-page-master master-name="all"
          page-height="11.5in"    margin-top="1in"          margin-left="1in"
          page-width="8.5in"      margin-bottom="0.75in"    margin-right="1in">
       <fo:region-body margin-top="0in" margin-bottom="0in"/>
       <fo:region-before extent="0in"/>
       <fo:region-after extent="0.5in"/>
     </fo:simple-page-master>
   </fo:layout-master-set>

   <fo:page-sequence master-reference="all">

   <fo:flow flow-name="xsl-region-body">
     <!-- article title -->
     <fo:block
          font-size="18pt"   font-family="sans-serif"   line-height="24pt"
          color="white"      text-align="center"        padding-top="0pt"
          background-color="black"   space-after.optimum="24pt">
       <xsl:value-of select="title" />
     </fo:block>

     <!-- process all sections -->
     <xsl:apply-templates select="section"/>

     </fo:flow>
   </fo:page-sequence>

 </fo:root>
 </xsl:template>

 <!-- process a section -->
 <xsl:template match="section">
   <!-- section title -->
   <fo:block
         font-size="14pt"     font-family="sans-serif"    font-weight="bold"
```

```
127        line-height="20pt"   space-after.optimum="12pt"  padding-top="0pt">
128    <xsl:value-of select="title"/>
129   </fo:block>
130
131   <!-- process a Glossary section -->
132   <xsl:if test="title='Glossary'">
133    <xsl:apply-templates select="entry"/>
134   </xsl:if>
135   </xsl:template>
136
137   <!-- process a glossary entry -->
138   <xsl:template match="entry">
139   <fo:block
140        text-align="start"    font-size="12pt"     text-indent="-0.5in"
141        margin-left="0.5in"   font-family="Times"  space-after.optimum="12pt">
142     <xsl:apply-templates select="term"/>
143   </fo:block>
144   </xsl:template>
145
146   <!-- process a term -->
147   <xsl:template match="term">
148   <fo:inline font-weight="bold" >
149    <xsl:value-of select="."/>
150    <xsl:apply-templates select="../definition"/>
151   </fo:inline>
152   </xsl:template>
153
154   <!-- process a definition -->
155   <xsl:template match="definition">
156   <fo:inline font-weight="normal">
157    - <xsl:value-of select="."/>
158   </fo:inline>
159   </xsl:template>
160
161   </xsl:stylesheet>
```

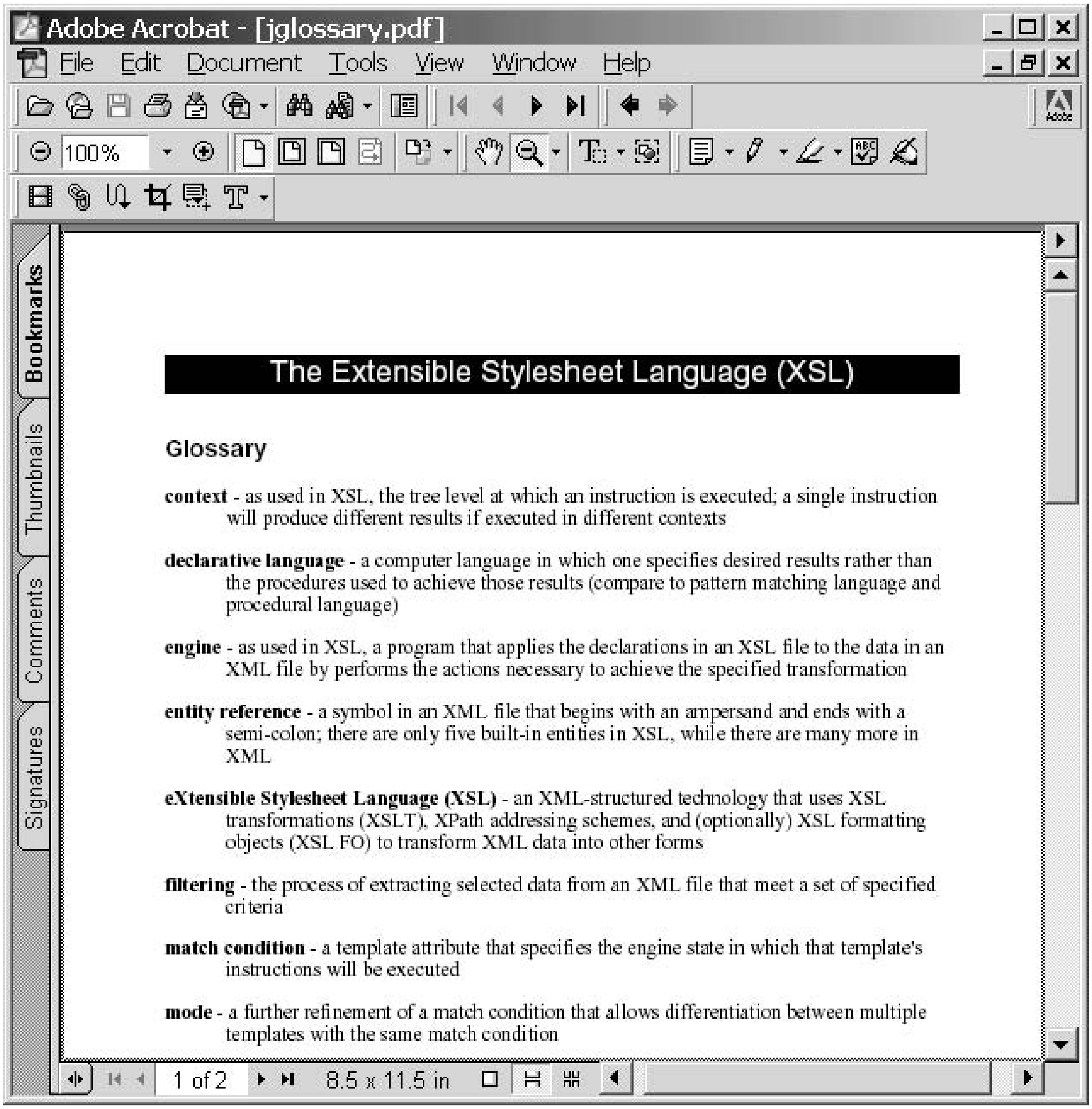
`master-set` element (lines 92–100). The *page sequence* begins at line 102. It starts a *flow* at line 104 and creates the first *block* at lines 106–111. Note, however, the one major difference between this example and the previous example: line 110 does not contain static text. Rather, line 110 inserts the text contained in the `title` child element of the XML `article` element using an `xsl:value-of` instruction. Line 114 then descends into the XML tree by applying the appropriate template that matches all the `section` child elements.

That template begins at line 123. It creates a new block at lines 125–129. Here again, at line 128, we use an `xsl:value-of` instruction to pull the text content to be rendered from the XML file rather than hard-coding it as static text in the XSL file. Note, however, that the `title` element here refers to a child of the XML `section` element, which is not the same as the `title` referred to at line 111. Refer back to lines 9 and 11 of the XML file in Listing 12 to see that `title` elements appear at two different levels. The recursive descent structure of the XSL processing engine ensures that we are referred to the right element at each level.

Next, controlled by the `xsl:if` instruction at line 132 to make sure that the section we're currently processing is a glossary, the `xsl:apply-templates` instruction at line 133 initiates processing of all glossary entries. Here's where the real fun begins!

Each `entry` element in the XML file has two child elements: `term` and `definition` (see lines 13 and 14 in Listing 12). Before displaying each term, the `entry` template begins a *block* (line 139 in Listing 13), which in this case is essentially a paragraph. Thus a line break occurs and the formatting instructions for that block are applied. All of the blocks in this layout have `space-after.optimum` set to 12 points (1/6 of an inch), essentially leaving a blank line between glossary entries. Line 142 then initiates processing of the `entry` element's `term` child element.

In the `term` template we see an `fo:inline` tag at line 148. This tag begins the definition of a section of text similar to a block, but `inline` indicates that no line break is to occur. The `fo:inline` tag is analogous to an HTML `<span>` tag: It delineates a section that has its own formatting but that is not separated from its surrounding elements by white space. Here the font weight is set to



The Extensible Stylesheet Language (XSL)

Glossary

context - as used in XSL, the tree level at which an instruction is executed; a single instruction will produce different results if executed in different contexts

declarative language - a computer language in which one specifies desired results rather than the procedures used to achieve those results (compare to pattern matching language and procedural language)

engine - as used in XSL, a program that applies the declarations in an XSL file to the data in an XML file by performs the actions necessary to achieve the specified transformation

entity reference - a symbol in an XML file that begins with an ampersand and ends with a semi-colon; there are only five built-in entities in XSL, while there are many more in XML

eXtensible Stylesheet Language (XSL) - an XML-structured technology that uses XSL transformations (XSLT), XPath addressing schemes, and (optionally) XSL formatting objects (XSL FO) to transform XML data into other forms

filtering - the process of extracting selected data from an XML file that meet a set of specified criteria

match condition - a template attribute that specifies the engine state in which that template's instructions will be executed

mode - a further refinement of a match condition that allows differentiation between multiple templates with the same match condition

Figure 16: Glossary transformed to PDF format using XSL-FO.

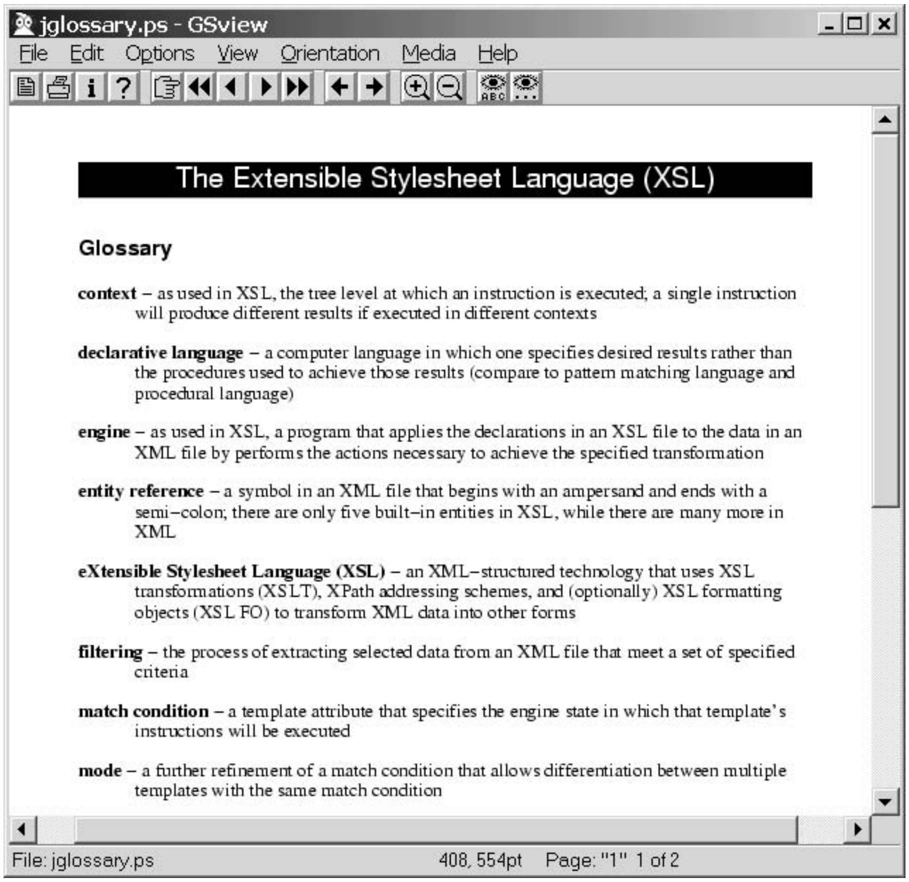
bold, and then the xsl:value-of instruction is used to insert the text of the current XML term element.

Line 150 continues processing by applying the definition template to the current term's sibling definition element. In that template (lines 155–159), the fo:inline tag is used to return the font weight to normal (not bold), and then a hyphen is rendered followed by the content of the XML definition element.

In addition to the finely tuned formatting control provided by XSL-FO tags and attributes, XSL-FO engines are now beginning to appear that can render those formats in a variety of ways. Figures 16–18 show the glossary excerpt we have been discussing rendered in PDF, PostScript, and Java AWT formats, respectively. All of these renderings were created from the same XML and XSL files using the FOP processing engine. To produce the different outputs, only the output specification parameter passed to the FOP program was changed.

WHERE TO GO FROM HERE

Hundreds—if not thousands—of pages would of course be needed to cover all the features of XSL and its XSLT, XPath, and XSL-FO subcomponents, but hopefully this chapter has given you enough information to grasp the essence of these powerful Web technologies. The References section provides pointers to further reference material and Web sites where you can not only learn about XSL, but also download the software discussed in this chapter to use XSL on your own systems.


jglossary.ps - GSview
File Edit Options View Orientation Media Help
File: jglossary.ps
408, 554pt Page: "1" 1 of 2

The Extensible Stylesheet Language (XSL)

Glossary

context – as used in XSL, the tree level at which an instruction is executed; a single instruction will produce different results if executed in different contexts

declarative language – a computer language in which one specifies desired results rather than the procedures used to achieve those results (compare to pattern matching language and procedural language)

engine – as used in XSL, a program that applies the declarations in an XSL file to the data in an XML file by performs the actions necessary to achieve the specified transformation

entity reference – a symbol in an XML file that begins with an ampersand and ends with a semi–colon; there are only five built–in entities in XSL, while there are many more in XML

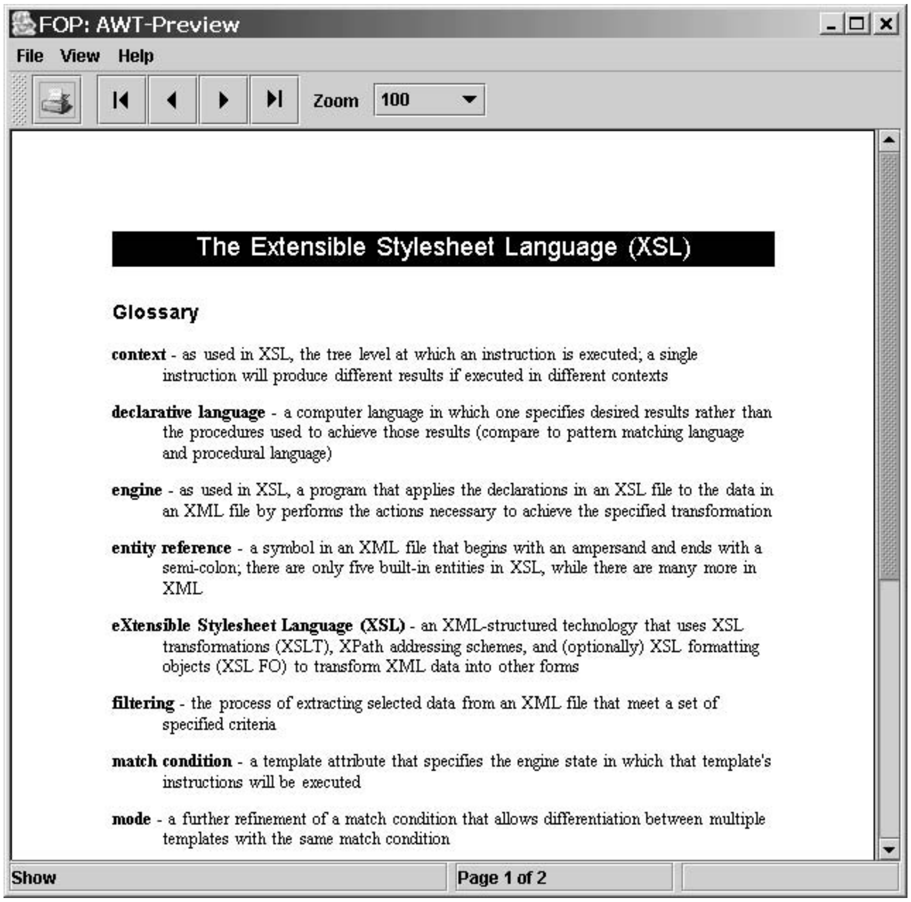
eXtensible Stylesheet Language (XSL) – an XML–structured technology that uses XSL transformations (XSLT), XPath addressing schemes, and (optionally) XSL formatting objects (XSL FO) to transform XML data into other forms

filtering – the process of extracting selected data from an XML file that meet a set of specified criteria

match condition – a template attribute that specifies the engine state in which that template's instructions will be executed

mode – a further refinement of a match condition that allows differentiation between multiple templates with the same match condition

Figure 17: Glossary transformed to PostScript format using XSL-FO.



The Extensible Stylesheet Language (XSL)

Glossary

context - as used in XSL, the tree level at which an instruction is executed; a single instruction will produce different results if executed in different contexts

declarative language - a computer language in which one specifies desired results rather than the procedures used to achieve those results (compare to pattern matching language and procedural language)

engine - as used in XSL, a program that applies the declarations in an XSL file to the data in an XML file by performs the actions necessary to achieve the specified transformation

entity reference - a symbol in an XML file that begins with an ampersand and ends with a semi-colon; there are only five built-in entities in XSL, while there are many more in XML

eXtensible Stylesheet Language (XSL) - an XML-structured technology that uses XSL transformations (XSLT), XPath addressing schemes, and (optionally) XSL formatting objects (XSL FO) to transform XML data into other forms

filtering - the process of extracting selected data from an XML file that meet a set of specified criteria

match condition - a template attribute that specifies the engine state in which that template's instructions will be executed

mode - a further refinement of a match condition that allows differentiation between multiple templates with the same match condition

Figure 18: Glossary transformed to Java AWT format using XSL-FO.

APPENDIX

Listing A1: XSL code to generate the output in Figure 2 from the XML in Figure 1.

```
<?xml version='1.0'?>
<!--
  File: cats7e.xsl
  updated by JMH on July 17, 2002 at 12:19 PM
-->
<xsl:stylesheet xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
                version="1.0">

<xsl:template match="/">
  <html>
  <body>

  <h2>Mary's Cats</h2>

  <table border="1" cellpadding="2" cellspacing="0">
    <thead>
      <td align="center"><b>Id</b></td>
      <td align="center"><b>Name</b></td>
      <td align="center"><b>Sex</b></td>
      <td align="center"><b>Breed</b></td>
      <td align="center"><b>Dob</b></td>
      <td align="center"><b>Color</b></td>
      <td align="center"><b>Kittens</b></td>
    </thead>
    <tbody>
      <xsl:for-each select="MyCats/Cat">
        <xsl:sort select="./text()" order="ascending"/>
        <tr>
          <td><xsl:value-of select="@id"/></td>
          <td><xsl:value-of select="./text()"/></td>
          <td align="center">
            <xsl:value-of select="@sex"/>
          </td>
          <td><xsl:value-of select="Info/Breed"/></td>
          <td><xsl:value-of select="Info/DOB"/></td>
          <td><xsl:value-of select="Info/@color"/></td>
          <td align="center">
            <xsl:choose>
              <xsl:when test="Info/Kittens/@number">
                <xsl:value-of select="Info/Kittens/@number"/>
              </xsl:when>
              <xsl:otherwise>
                --
              </xsl:otherwise>
            </xsl:choose>
          </td>
        </tr>
      </xsl:for-each>
    </tbody>
  </table>

  </body>
  </html>
 </xsl:template>

 </xsl:stylesheet>
```

Listing A2: XSL code to generate the output in Figure 3 from the XML in Figure 1.

```
<?xml version='1.0'?>
<!--
  File: cats7e-3.xsl
  updated by JMH on July 17, 2002 at 12:19 PM
-->
<xsl:stylesheet xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
                version="1.0">

<xsl:template match="/">
  <html>
  <body>

  <h2>Mary's Cats</h2>

   <table border="1" cellpadding="2" cellspacing="0">
    <thead>
      <td align="center"><b>Id</b></td>
      <td align="center"><b>Name</b></td>
      <td align="center"><b>Sex</b></td>
      <td align="center"><b>Breed</b></td>
      <td align="center"><b>Dob</b></td>
      <td align="center"><b>Color</b></td>
      <td align="center"><b>Kittens</b></td>
    </thead>
    <tbody>
      <xsl:apply-templates select="MyCats/Cat">
        <xsl:sort select="Info/DOB" order="ascending"/>
      </xsl:apply-templates>
    </tbody>
   </table>

   </body>
   </html>
 </xsl:template>

 <xsl:template match="Cat">
   <tr>
    <td><xsl:value-of select="@id"/></td>
    <td><xsl:value-of select="./text()"/></td>
    <td align="center">
      <xsl:value-of select="@sex"/>
    </td>
    <td><xsl:value-of select="Info/Breed"/></td>
    <td><xsl:value-of select="Info/DOB"/></td>
    <td><xsl:value-of select="Info/@color"/></td>
    <td align="center">
      <xsl:choose>
        <xsl:when test="Info/Kittens/@number">
         <xsl:value-of select="Info/Kittens/@number"/>
        </xsl:when>
        <xsl:otherwise>
          --
        </xsl:otherwise>
      </xsl:choose>
    </td>
   </tr>
 </xsl:template>

 </xsl:stylesheet>
```

Listing A3: XSL code to generate the output in Figure 4 from the XML in Figure 1.

```
<?xml version='1.0'?>
<!--
  File: cats7e-4.xsl
  updated by JMH on July 17, 2002 at 01:48 PM
-->
<xsl:stylesheet xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
                version="1.0">

<xsl:template match="/">
  <html>
  <body>

  <h2>Mary's Cats</h2>

  <table border="1" cellpadding="2" cellspacing="0">
    <thead>
      <td align="center"><b>Id</b></td>
      <td align="center"><b>Name</b></td>
      <td align="center"><b>Kittens</b></td>
    </thead>
    <tbody>
      <xsl:apply-templates select="MyCats/Cat[Info/Kittens/@number]"/>
    </tbody>
   </table>

   </body>
   </html>
 </xsl:template>

 <xsl:template match="Cat">
   <tr>
    <td><xsl:value-of select="@id"/></td>
    <td><xsl:value-of select="./text()"/></td>
    <td align="center">
      <xsl:value-of select="Info/Kittens/@number"/>
    </td>
   </tr>
 </xsl:template>

 </xsl:stylesheet>
```

Listing A4: JavaServer Page code to apply an XSL file to an XML file using the Apache Xalan-Java XSL engine.

```
<html>
<!--
  Code to Apply an XSL File to an XML File on the Server Side
  adapted from xalan-j_2_3_1\samples\SimpleTransform\SimpleTransform.java
  download Xalan-Java from http://xml.apache.org/xalan-j free of charge
  updated by JMH on July 16, 2002 at 07:23 PM
-->
<head>
  <%@ page import="java.io.*"%> <!-- for StringWriter -->
  <%@ page import="javax.xml.transform.*, javax.xml.transform.stream.*" %>

 <%!
   /** Apply an XSL file to an XML file and return the results as a String.
     * @param strXMLfile String containing full URL to the XML file
   * @param strXSLfile String containing full URL to the XSL file
   * @return String containing the output of the transformation
   */
  String ApplyXSL( String strXMLfile, String strXSLfile )
  {
    // StringWriter is a child of java.io.Writer and can therefore be
    // used as an argument to a StreamResult constructor, which is
    // required by Transformer.transform().
    StringWriter swResult = new StringWriter() ;

   try {
    // Use the static TransformerFactory.newInstance() method to instantiate
    // a TransformerFactory. The javax.xml.transform.TransformerFactory
    // system property setting determines the actual class to instantiate --
    // org.apache.xalan.transformer.TransformerImpl.

    TransformerFactory tFactory = TransformerFactory.newInstance();

    // Use the TransformerFactory to instantiate a Transformer that will work
    // with the stylesheet you specify. This method call also processes the
    // stylesheet into a compiled Templates object.

    Transformer transformer =
        tFactory.newTransformer( new StreamSource( strXSLfile ) );

    // Use the Transformer to apply the associated Templates object to an XML
    // document and output the result.

    transformer.transform( new StreamSource( strXMLfile ),
                           new StreamResult( swResult ) );

    // Return the result.
    return swResult.toString();

   } catch ( TransformerConfigurationException tfce ) {
     return tfce.toString() ;

   } catch ( TransformerException tfe ) {
     return tfe.toString() ;
   }
  }
%>
</head>

<body>
<%
  // This code assumes that the XML and XSL files reside in the same directory
  // as this JSP. Those file names are hard-code here, but they could be passed
  // as parameters from an HTML form or via some other such technique.
  String strXMLfilename = "hello.xml" ;  // replace with your XML file name
  String strXSLfilename = "hello.xsl" ;  // replace with your XSL file name

  // We need to construct a full path to the XML and XSL files, including the
  // "http://" protocol specification, server name, and server port number (if
  // it's not the default of 80). The following code accomplishes this.
  String strFullPath = "http://" + request.getServerName() ;
  if ( request.getServerPort() != 80 )
    strFullPath += ":" + request.getServerPort() ;

  // We then add the full path to this JSP and finally strip off this JSP's file
  // name and extension that follow the last forward slash (/).
  strFullPath += request.getRequestURI() ;
  strFullPath = strFullPath.substring( 0, strFullPath.lastIndexOf( "/" ) + 1 );

  // We append the XML and XSL file names to the full path to pass them to our
  // ApplyXSL method, which applies the XSL file to the XML file and returns the
  // result as a string. Printing that String shows the result in the browser.
  out.println( ApplyXSL( strFullPath + strXMLfilename,
                         strFullPath  + strXSLfilename ) );
%>
</body>
</html>
```

Listing A5: Java Servlet to apply an XSL file to an XML file using the Apache Xalan-Java XSL engine.

```
/*
 * ApplyXSLServlet.java
 *
 * Created using Forte for Java 4, Community Edition, on July 17, 2002, 10:30 AM
 */

 package InternetEncyclopedia.XSLDemos ;

 import javax.servlet.*;
 import javax.servlet.http.*;

 import java.io.* ;    // for StringWriter and PrintWriter
 import javax.xml.transform.* ;
   // for Transformer, TransformerFactory, TransformerException,
   // and TransformerConfigurationException
 import javax.xml.transform.stream.* ;
   // for StreamSource and StreamResult

 /**
  * This servelt applies an XSL file to an XML file and displays the results.
  * @author Jesse M. Heines
  * @version 1.0
  */
 public class ApplyXSLServlet extends HttpServlet
 {
   /** Apply an XSL file to an XML file and return the results as a String.
    *  @param  strXMLfile String containing full URL to the XML file
    *  @param  strXSLfile String containing full URL to the XSL file
    *  @return String containing the output of the transformation
    */
   private String ApplyXSL( String strXMLfile, String strXSLfile )
   {
   // StringWriter is a child of java.io.Writer and can therefore be
   // used as an argument to a StreamResult constructor, which is
   // required by Transformer.transform().
   StringWriter swResult=new StringWriter();

   try {
     // Use the static TransformerFactory.newInstance() method to instantiate
     // a TransformerFactory. The javax.xml.transform.TransformerFactory
     // system property setting determines the actual class to instantiate --
     // org.apache.xalan.transformer.TransformerImpl.

    TransformerFactory tFactory = TransformerFactory.newInstance();

    // Use the TransformerFactory to instantiate a Transformer that will work
    // with the stylesheet you specify. This method call also processes the
    // stylesheet into a compiled Templates object.

    Transformer transformer =
        tFactory.newTransformer( new StreamSource( strXSLfile ) );

    // Use the Transformer to apply the associated Templates object to an XML
    // document (foo.xml) and write the output to a file (foo.out).

    transformer.transform( new StreamSource( strXMLfile ),
                           new StreamResult( swResult ) );

    // Return the result.
    return swResult.toString() ;

    } catch ( TransformerConfigurationException tfce ) {
      return tfce.toString() ;

    } catch ( TransformerException tfe ) {
      return tfe.toString() ;
    }
  }

  /** Display an error message for a missing field.
   *  @param strErrorMsg String containing error message to show
   *  @return String containing the output of the transformation
   */
  private void ShowErrorMessage( PrintWriter out, String strErrorMsg )
  {
    out.println( "<p><font color='red'>" + strErrorMsg + "</font></p>" );
    out.println( "<p>Please press your browser's BACK button and try again.</p>" );
  }

  /** Initializes the servlet. (supplied by Forte for Java 4)
   */
  public void init(ServletConfig config) throws ServletException
  {
    super.init(config);
  }

  /** Destroys the servlet.  (supplied by Forte)
   */
  public void destroy()
  {}

  /** Processes both HTTP <code>GET</code> and <code>POST</code> methods.
    *  @param  request   servlet request
    *  @param  response  servlet response
    */
  protected void processRequest(
      HttpServletRequest request, HttpServletResponse response )
      throws ServletException, java.io.IOException
  {
     // This code assumes that the XML and XSL files reside in the same directory.
     String strXMLpath = request.getParameter( "XMLpath" );
     // XML file name passed from an HTML form
     String strXMLfilename = request.getParameter( "XMLfilename" ) ;
   // XSL file name passed from an HTML form
   String strXSLfilename = request.getParameter( "XSLfilename" );

   response.setContentType("text/html");
   java.io.PrintWriter out = response.getWriter();

   out.println( "<html>" ) ;
   out.println( "<head>" ) ;
   out.println( " <title>Apply XSL Servlet</title>" ) ;
   out.println( "</head>" ) ;

   out.println( "<body>" ) ;

   if ( ( strXMLpath == null ) ||  strXMLpath.equals( "" ) ) {
     ShowErrorMessage( out, "No path supplied for XML and XSL files." ) ;
   } else if ( ( strXMLfilename == null ) || strXMLfilename.equals( "" ) ) {
     ShowErrorMessage( out, "No name supplied for your XML file." ) ;
   } else if ( ( strXSLfilename == null ) || strXSLfilename.equals( "" ) ) {
     ShowErrorMessage( out, "No name supplied for your XSL file." ) ;
   } else {
     out.println( ApplyXSL( strXMLpath + strXMLfilename,
                            strXMLpath + strXSLfilename ) );
   }

     out.println( "</body>" ) ;
     out.println( "</html>" ) ;

     out.close();
   }

   /** Handles the HTTP <code>GET</code> method. (supplied by Forte for Java 4)
    *  @param request servlet request
    *  @param response servlet response
    */
  protected void doGet(HttpServletRequest request, HttpServletResponse response)
  throws ServletException, java.io.IOException
  {
    processRequest(request, response);
  }

  /** Handles the HTTP <code>POST</code> method. (supplied by Forte for Java 4)
   *  @param request servlet request
   *  @param response servlet response
   */
  protected void doPost(HttpServletRequest request, HttpServletResponse response)
  throws ServletException, java.io.IOException
  {
   processRequest(request, response);
  }

 /** Returns a short description of the servlet. (supplied by Forte for Java 4)
  */
 public String getServletInfo()
 {
   return "This small servlet applies an XSL file to an XML file and displays the " +
          "results.";
  }
}
```

Listing A6: HTML form to supply parameters to the Apply XSL Java Servlet.

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN">
<!--
  ApplyXSLForm.htm: to provide parameters for the ApplyXSLForm servlet
  Jesse M. Heines, UMass Lowell Computer Science, heines@cs.uml.edu
  updated by JMH on July 17, 2002 at 10:54 AM
-->
<html>
  <head>
    <title>Apply XSL Form</title>
    <script type="text/javascript">
      function init( frm ) // called when body is loaded to initialize form
      {
        // set default path to the location of this form
        var strPathName = "http://" + location.host ;
        strPathName += location.pathname.substring(
                          0, location.pathname.lastIndexOf( "/" ) + 1 ) ;
        frm.XMLpath.value = strPathName ;   //set default path in first form field
        frm.XMLfilename.focus() ;           //set focus to second form field
      }
    </script>
  </head>
  <body onload="init( frm )">
    <p>Please fill in the fields below to specify the parameters needed to
    apply an XSL file to an XML file.</p>
    <p><i>Notes:</i></p>
    <ul>
      <li>This form expects your XML and XSL files to be in the same directory.</li>
      <li>You may edit the full path field, but your entry must be of the form shown
          initially.</li>
    </ul>
    <br />
    <form name="frm" action="/servlet/InternetEncyclopedia.XSLDemos.ApplyXSLServlet">
      <table name="tbl">
        <tr>
          <td align="right">Path to XML and XSL files:</td>
          <td><input name="XMLpath" size="60"/></td>
        </tr>
        <tr>
          <td align="right">XML file name:</td>
          <td><input name="XMLfilename" size="60" /></td>
        </tr>
        <tr>
          <td align="right">XSL file name:</td>
          <td><input name="XSLfilename" size="60" /></td>
        </tr>
        <tr>
          <td align="right"></td>
          <td align="right">
            <input type="submit" value="Submit Entries"/>
            <input type="reset" value="Clear Fields" />
          </td>
        </tr>
      </table>
    </form>
  </body>
</html>
%
```

GLOSSARY

Context As used in XSL, the tree level at which an instruction is executed; a single instruction will produce different results if executed in different contexts.

Declarative language A computer language in which one specifies desired results rather than the procedures used to achieve those results (compare to *pattern matching language* and *procedural language*).

Engine As used in XSL, a program that applies the declarations in an XSL file to the data in an XML file by performing the actions necessary to achieve the specified transformation.

Entity reference A symbol in an XML file that begins with an ampersand and ends with a semi-colon; there are only five built-in entities in XSL, while there are many more in XML.

Extensible stylesheet language (XSL) An XML-structured technology that uses XSL transformations (XSLT), XPath addressing schemes, and (optionally) XSL formatting objects (XSL-FO) to transform XML data into other forms.

Filtering The process of extracting selected data from an XML file that meet a set of specified criteria.

Match condition A template attribute that specifies the engine state in which that template's instructions will be executed.

Mode A further refinement of a match condition that allows differentiation between multiple templates with the same match condition.

Namespace A collection of externally predefined words that are used in an XML document as element types and attribute names; the namespace is specified as a uniform resource identifier, which for XSL is `http://www.w3.org/1999/XSL/Transform`.

Namespace prefix A user-chosen name used to refer to the namespace to which an XML tag name belongs.

Node As used in XSL, an object in an XML tree that has zero or one parent and zero or more children; the abstract representation of an XML tag or attribute or text content.

Pattern matching language A language for specifying constraints ("patterns") under which actions are to take place (compare to *declarative language* and *procedural language*).

Procedural language A computer language in which one specifies the precise steps required to achieve a desired result (compare to *declarative language* and *pattern matching language*).

Processing instruction A command in an XML file that specifies file attributes such as the version number or stylesheet link; a processing instruction is not part of the XML data.

Recursive descent The process of visiting each node in a tree or subtree by recursively tracing the links to each node's child to traverse the entire tree structure.

Root element The "top" element of a tree that has no parent node.

Style An attribute of a visible element such as its color, size, font type, etc., or any other characteristic appropriate for some other type of element (compare to *transformation*).

Stylesheet A collection of specifications that define how information is presented; stylesheets can define how text data is rendered on a screen, how words are pronounced, etc.

Template A collection of instructions that specify the desired result of a transformation in a specific context and (optionally) mode.

Transformation The process of generating a new form of data from another form (compare to *style*).

XPath The expression language used to refer to collections of XML nodes for processing by XSLT.

XSL formatting objects (XSL-FO) An XML namespace that defines a language for transforming XML data into a formatted presentation.

XSL transformations (XSLT) The language used by XSL to transform XML data into other forms.

CROSS REFERENCES

See *Cascading Style Sheets (CSS); Extensible Markup Language (XML).*

REFERENCES

Anderson, R., Birbeck, M., Kay, M., Livingstone, S., Loesgen, B., Martin, D., Mohr, S., Ozu, N., Peat, B., Pinnock, J., Stark, P., & Williams, K. (2000). *Professional XML*. Birmingham, UK: Wrox Press.

Apache Software Foundation. (2002). Xalan-Java. Retrieved July 16, 2002, from http://xml.apache.org/xalan-j

Cagle, K. (2000a). Transform your data with XSL. *XML Magazine 1*(1), 76–80.

Cagle, K. (2000b). ArchitectureX: Designing for XML. *XML Magazine 1*(2), 22–28.

Kay, M. (2000). *XSLT programmer's reference* (1st ed.). ISBN 1861005067. Birmingham, UK: Wrox Press.

Microsoft Corporation (2002). *Microsoft XML core services (MSXML) 4.0—XSLT reference*. Retrieved July 16, 2002, from http://msdn.microsoft.com/library/default.asp?url = /library/en-us/xmlsdk/htm/xsl_ref_overview_1vad.asp

Oracle Corporation (1999, updated 2001). Using XML in Oracle database applications: Exchanging business data among applications. Retrieved white paper on July 16, 2002, from http://technet.oracle.com/tech/xml/info/htdocs/otnwp/xml_data_exchange.htm

World Wide Web Consortium (1997). Date and time formats. Retrieved "Note" on April 18, 2002, from http://www.w3.org/TR/NOTE-datetime.html

World Wide Web Consortium (1999). Namespaces in XML. Retrieved "Recommendation" on July 16, 2002, from http://www.w3.org/TR/REC-xml-names

World Wide Web Consortium (1999). XSL Transformations (XSLT) (Version 1.0). Retrieved "Recommendation" on July 16, 2002, from http://www.w3.org/TR/xslt

World Wide Web Consortium (2001). Extensible Stylesheet Language (XSL) (Version 1.0). Retrieved "Recommendation" on July 16, 2002, from http://www.w3.org/TR/2001/REC-xsl-20011015

World Wide Web Consortium. (2002). XSL Transformations (XSLT)(Version 2.0). Retrieved "Working Draft" on July 16, 2002, from http://www.w3.org/TR/xslt20

FURTHER READING

Reference Web sites (*in alphabetical order*)

http: // msdn.microsoft. com/ library/ default. asp? url = /library/en-us/xmlsdk30/htm/xmrefxsltreference.asp—comprehensive and very well organized reference material on all aspects of XSLT and XPath.

http://www.dpawson.co.uk/xsl—extremely comprehensive list of XSL frequently asked questions with extensive answers and examples.

http://www.jenitennison.com/xslt/index.html—wonderful tutorials and documentation on many advanced applications of XSLT and XPath.

http://www.mulberrytech.com/xsl—wealth of XSL reference material including wonderful "quick reference" sheets for XML, XSLT, and XPath; also the parent of the phenomenally active XSL-List Open Forum at http://www.mulberrytech.com/xsl/xsl-list/index.html

http://www.netcrucible.com/xslt/msxml-faq.htm—*unofficial* answers to MSXML XSLT frequently asked questions.

http://www.oasis-open.org/cover/xsl.html—a comprehensive list of online references for XSL, XSLT, XPath, and related standards.

http://www.w3.org/Style/XSL—World Wide Web Consortium home page for all XSL-related documents.

http://www.w3.org/TR/xpath—the official World Wide Web Consortium's recommendation for the XML Path Language, XPath.

http://www.w3.org/TR/xslt—the official World Wide Web Consortium's recommendation for XSL Transformations, XSLT.

http://xml.apache.org/fop—home page for the Apache Group's FOP Formatting Objects Processor, a print formatter driven by XSL formatting objects.

Extranets

Stephen W. Thorpe, *Neumann College*

INTRODUCTION

The Internet is a wide-area network that consists of thousands of organizational networks worldwide. The Internet became popular in the 1990s with the advent of the “World Wide Web.” The Web provides information stored on servers that are connected to the Internet. Web browsers, such as Netscape Navigator and Microsoft Internet Explorer, provide interfaces to interpret and display the Web pages. Documents accessible on the Web contain hyperlinks to move from one document to another. The benefits of the Internet include the ability to discover information more quickly and in larger volumes as well as increased opportunities for communication with individuals and groups. It is hard to imagine an organization today that does not have access to the Internet or does not maintain its own organizational Web site. In the early days of the Internet, an organizational presence on the Web provided a competitive advantage. Today, an organization’s presence on the Web is an essential part of doing business.

The benefits of Internet technology and Web browsing led to the creation of organizational networks that provide restricted access for internal employees. The internal networks became known as “intranets,” private enterprise networks that use the Internet and Web technologies for information gathering and distribution within the organization. By converting old applications or writing new applications for use on an intranet, organizations can eliminate dependence on a particular operating system/platform. The intranet has also become an effective communication vehicle to share information within the organization. The intranet provides easy access to internal information that can be published by departments within the organization. Other applications supported by the intranet include employee access to legacy systems, Web sites for human resources, payroll information, and even training programs. Intranets support collaborative processes between departments, such as scheduling, messaging, and discussion groups (Lloyd & Boyle, 1998). Intranets can also empower customer support center staff with intranet knowledge management systems that result in improved customer satisfaction (Phaltankar, 2000). Through the use of intranets, organizations have reduced costs while also increasing employee productivity.

The intranet is typically protected from the public Internet by a firewall, a device that acts as a gatekeeper between the organizational intranet and the outside Internet. The firewall permits internal employees to access the public Internet, but it prevents outside users (the public) from accessing the internal resources of the organization.

A natural extension of the intranet would allow selected external entities to access internal organizational resources. An “extranet” provides a collaborative network that uses Internet technologies to extend intranet applications and services to selected external organizations, thereby creating a community of interest beyond the walls of the organization. In addition, providing remote access to the intranet permits employees who are off site to access intranet services; for example, employees working from home could access intranet resources through the deployment of an extranet. The extranet is typically not available to the general public but rather it is limited to “strategic partners.” These strategic partners may include suppliers, vendors, customers, or other organizations involved in collaborative research and development. Bringing suppliers, partners, and even customers into the information loop is critical to developing a company’s quick response and strategic movement as it adapts to an evolving market environment (Baker, 2000). Figure 1 depicts an organizational extranet.

STRATEGIC USES OF EXTRANETS

The role and importance of extranets appear to be increasing. In a 1999 study by Forrester Research of 50 large manufacturers who had implemented extranets, 80% of the interviewees said they expected to extend extranet access to all of their business partners within two years (Orlov, 1999). While over two-thirds of those surveyed were already providing marketing and product specifications through their extranets, virtually all of the executives interviewed expected to expand their extranets to include online sales and sharing of inventory information. The manufacturers also anticipated a 32% reduction

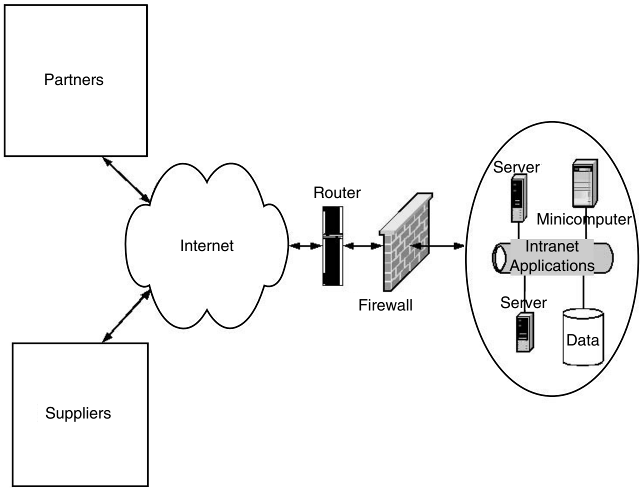


Figure 1: Organizational extranet.

in support costs and a 17% increase in sales from their extranet applications.

Extranets provide several strategic advantages for organizations to extend their intranet applications to their business partners: business-to-business e-commerce, collaboration for research and development, and internal efficiency. Throughout the following section, several examples of extranets will be described to illustrate the application of extranets. The illustrations are intentionally drawn from a variety of industries to demonstrate the ubiquity of extranet deployment in business.

Business-to-Business E-commerce

Business-to-business (B2B) e-commerce has taken off in recent years through the maturity of the Internet and the Web. Extranets provide a means to share existing intranet content for real-time communication between business partners. The extranet also provides external organizations with access to internal databases and legacy systems without necessarily connecting to the full intranet of the organization. Among the advantages of e-commerce, the extranet can be an effective tool in customer relationship management, supply chain management, and the electronic exchange of business transactions.

Customer Relationship Management (CRM)

Customer relationship management involves integrating the customer service processes with a business, such as sales, marketing, customer service, and support. The idea behind CRM is to adapt business processes to a "customer-centric" business strategy. The deployment of an extranet achieves CRM strategies by providing customers with 24-hours-a-day, seven-days-a-week access to internal information without the involvement of business personnel. For example, extranets can support business-to-business transactions, such as sales, while also providing information resources for products and services. Several customer service components of an organization could potentially be deployed through an extranet.

In the health care industry, Humana has developed an extranet application for physicians, consumers, employers, and brokers that may well transform the company and establish it as a technology leader in health care (Lewis, 2001b). The extranet site, Emphesys.com, provides a customer-centric approach toward delivering information in real time for various organizational constituencies. The extranet offers Web portals for doctors, patients, and others to share real-time data as well as to access information from a medical database. The extranet includes links for doctors to access the National Institutes of Health and other research resources. In addition, physician offices will have online access to patient data ranging from claims status to prescription histories. For health care plan members, the extranet application provides information regarding Humana's various health plans and allows individuals to review their medical records and prescription data. The extranet application will permit members to enroll in the health care plan of their choice and provide expedited plan authorization. Efficiencies are gained through the extranet, which will provide a member's enrollment eligibility information to doctors within four hours, compared to two or more weeks that had been necessary before the extranet.

Even small organizations can deploy extranets to improve customer access to internal information. The law offices of Dinsmore & Shohl established a subsidiary organization, DinsNet Technologies, to offer electronic services for clients and lawyers (Baker, 2000). Because of a number of cases that were document intensive, the law office began scanning the documents to make them available in electronic format. The DinsNet case management services extranet allows clients to access documents as well as billing statements and provides for immediate client/attorney interaction and document sharing. The extranet provides enhanced flexibility for communication with clients who are not physically near the law offices.

Supply Chain Management

Supply chain management involves the integration of business processes and technology to optimize information and product flows among the business partners within the supply chain, from suppliers to internal logistics, distribution, and customers. By providing suppliers with access to organization inventory systems through an extranet, the organization can implement "just-in-time" inventory replenishment policies. This provides a cost savings since the inventory is not maintained by the organization, but rather by the supplier. Extranets also provide the possibility to outsource the function of inventory management to suppliers.

For example, Coors Brewing Company, which has always been challenged to forecast product demand, has deployed an extranet application that allows over 1,000 distributors to order beer, access promotional information that might impact sales, and communicate with other distributors (Moozakis, 2001). The extranet allows distributors to populate a database with forecasts that Coors can then aggregate to make its own demand projections. The extranet provides basic ordering applications and will provide market analysis for distributors to understand better their regional demand. Other features planned for the extranet include inventory management and shipping applications that will connect to Coors' enterprise resource planning system.

Cisco Systems, a networking equipment maker, has deployed an extranet that links their component suppliers, distributors, and manufacturers to form a virtual, just-in-time supply chain (Koch, 2000). When a customer order arrives, typically through Cisco's Web site, a series of electronic messages are sent through the extranet to their business partners to produce the requested equipment. Contract manufacturers are able to access the supply chain process through Cisco's extranet that links to their manufacturing execution systems. On the other hand, Cisco has access to the contractor's assembly line and, when the equipment is ready, a connection is established for Cisco to complete product testing through their automated testing software. When testing is completed satisfactorily, Cisco releases the customer name and shipping address so the contractor can complete distribution to the customer. Through this innovative use of extranet applications, Cisco reduces or eliminates entirely the need for inventory space and paper invoices.

Electronic Data Interchange (EDI)

Extranets provide an economical method for the exchange of business transactions between organizations. To gain efficiencies in business-to-business transactions, large organizations adopted strategies for electronic data interchange to conduct business transactions through private, long-distance networks. Electronic data interchange is a computer-to-computer exchange of business information in a standard electronic format between business partners. These networks, however, required propriety protocols and were expensive to implement. The high cost of investing in private networks and the hardware and software for EDI was prohibitive for smaller organizations that could not realize the benefits of this technology.

Unlike traditional EDI systems that required private networking solutions, the extranet uses the public Internet to exchange transactional data. By eliminating the costs of private networking technologies, an extranet provides an economical alternative for medium and smaller companies to exchange transactional data with other businesses (Pfaffenberger, 1998; Salamone, 2000). Because the extranet uses the public Internet and TCP/IP protocols, it provides an open platform that is not limited to only participants who use the same protocols, Internet service providers, or operating systems (Anderson, 1997).

Collaboration with Business Partners

Communication up and down the supply chain with business partners is typically handled by telephone, fax, or e-mail. The extranet, however, provides the ability to get all parts of the business community working together and communicating in a way that is more timely and effective (Meyer, 2000). By putting employees, customers, and strategic partners in closer contact, extranets promote real-time collaborative efforts and expedite the time to market for products. Because everyone involved in a project can see proposed changes in real time, the extranet can also support many more iterations of the design process as compared to traditional techniques (Pfaffenberger, 1998).

Marshall Industries, an electronics distributor, developed an extranet that facilitates the flow of information to its hundreds of suppliers, systems integrators, and value-added resellers who can get point-of-sales reports at any time as well as real-time inventory access (Anderson, 1997). The extranet application, PartnerNet, provides both suppliers and customers with the ability to create customized information views from their corporate intranet. When new information is available, PartnerNet notifies the registered user.

One of the critical steps in designing an extranet is defining the differences and levels of access needed by business partners who might need to access the extranet. Marshall Industries recognized the differences among individual partners and does not provide equal access to their extranet for all members. A sales representative, for example, may be restricted to information regarding assigned accounts. A manager-level representative, however, may be provided access to regional-level sales activity for that manager's region, but the representative would not be provided with information from other regions.

The McCann Worldgroup is a marketing communications agency that uses collaboration software for internal and external communications to provide employees and clients with more effective ways to share work in progress (Maddox, 2001). The portal product, McWisdom, is both an intranet for internal communications and an extranet for client collaboration. By logging onto the secure area of the McCann Web site, employees as well as clients can review media plans, status reports on ad campaigns, and reports on campaign performance. The portal also provides clients with personalized views of projects in development.

Caterpillar, a manufacturer of heavy equipment, has developed extranets to facilitate collaboration between engineers and suppliers. The collaboration includes sharing and marking up design drawings and allowing suppliers to conduct their own simulations and testing. Through these efforts, Caterpiller has shrunk its product development cycle from as long as eight years to 39 months (Hills, 1998).

Internal Efficiencies

The use of an extranet can simplify workflow and create internal efficiencies. By allowing authorized employees to access supplier extranets, for example, organizations can simplify processes for ordering supplies and products (Ling & Yen, 2001). The Dell Computer Premier Dell.com Web site provides a useful example of how organizations can gain internal efficiencies through extranet applications. Litton PRC, a global systems integrator, began planning for an electronic purchasing system in 1999 (for more information, see http://www.dell.com/us/en/casestudies). Prior to this time, Litton PRC's procurement system was paper-based and required purchase requisitions from their worldwide offices to be reviewed by the main headquarters. On average, this paper-based system required 16.2 days to get the purchase requisition to the Purchasing office and up to as many as 17 approvals. Once the extranet application was implemented, the time required for Litton PRC to purchase Dell computers dropped to an average of 1.9 days from the time of creating a requisition order to delivery.

The use of extranets also provides internal efficiencies in the distribution of corporate information to external partners. Conoco, the nation's number five oil refiner, developed a partner-focused B2B Internet initiative to deliver invoices, credit card settlement, and electronic funds transfer information to its 5,000-plus wholesale and retail marketers (Lewis, 2001a). Prior to deploying the extranet, Conoco relied on faxes and other methods to distribute accounting data to marketers. The extranet application provides more timely feedback of information to their business partners.

The Ketchum extranet speeds up interoffice communications and provides an efficient method to help employees learn about new client businesses (Clark, 2001). Ketchum, a public relations firm, introduced MyKGN in March 2000, which allows employees to do research and work with customers online. In the past, research was done on a piecemeal basis. With the extranet, employees are able to research subject databases quickly and efficiently, while also working with clients in real time on such things as press releases and company logos. Ketchum has deployed over 400 collaborative client sites that are using MyKGN, and estimates that its extranet application saves 90 minutes a week per employee. By 2003, Ketchum may realize up to $5.6 million in cash flow and productivity benefits (Clark, 2001).

Organizations can realize internal efficiencies by enabling customers or suppliers to search internal business databases to find and order products, check pricing and availability, and find product documentation without the intervention of employees within the organization. Beyond the benefits of providing faster access to information for customers, extranet applications relieve some of the demands on customer support service personnel (Pfaffenberger, 1998). SunLife Financial, for example, recently deployed an extranet to attract insurance brokers to do business with them rather than their competitors. They hope the extranet will provide a competitive advantage while also holding down costs by allowing insurance brokers to obtain information without having to contact SunLife's call center. By making information available through the extranet, SunLife expects to be able to expand their business without having to increase call center resources (Messmer, 2000).

PLANNING AND IMPLEMENTING AN ORGANIZATIONAL EXTRANET

Planning and implementing an extranet require careful coordination between the organization and its business partners to identify the best business opportunities for an extranet, as well as an assessment of the potential risks to the organization once an extranet is implemented (Szuprowicz, 1998). Important issues to consider in building an extranet include defining the purpose and intended audience of the extranet, determining the return on investment, developing a content management strategy, and security. Of these issues, clearly security of internal information that will be shared over the public Internet is a critical concern. However, defining the purpose and applications for the extranet is also critical if the extranet is to provide the organization with a competitive advantage. In addition, outsourcing the development of the extranet is a possibility for those organizations that do not have the time or expertise to invest in design, implementation, and management of an extranet.

Defining the Purpose for an Extranet

Developing an extranet involves a shift in management philosophy since traditionally internal corporate information was intended to remain internal. Changing the management philosophy will likely meet with cultural barriers among employees who may not be eager to share internal data with outside organizations (Harreld, 2000). People who controlled information in the past may feel threatened by an extranet approach that eliminates information silos and potentially the power bases of some individuals. It is a management imperative to anticipate and resolve internal opposition in order for the development and implementation of an extranet to be successful; in fact, failing to recognize and address resistance is one reason for extranet failure (Pfaffenberger, 1998; Szuprowicz, 1998).

In deciding to implement an extranet, the organization must have a clear purpose in mind, which in part requires a determination of the target audience. With extranet applications, the target audience is typically an established business partner with whom the organization is looking to make doing business easier, faster, and more economical. Without a clear understanding of the purpose and audience for the extranet, content developers will be unable to create a satisfactory site.

While extranet design will be specific to the organization and its business opportunities, extranets tend to fall into three general categories: publishing, collaboration, or transactional (Pfaffenberger, 1998; Szuprowicz, 1998). Publishing extranets are intended to make internal documents available to external users. These documents might include technical documentation, training materials, specifications, or research data. Publishing extranets rely on servers and scripts to database applications for document retrieval. The primary focus of a publishing extranet is cost reduction by eliminating paper, mailing and faxing expenses, and even reducing customer service center expenses (Szuprowicz, 1998).

Collaboration extranets are designed to provide interaction between internal employees and external partners. The focus of collaboration extranets is to reduce the time to market of products by accelerating the product design cycle. The extranet examples cited earlier from the McCann Worldgroup and Ketchum facilitate communication between employees and clients in planning, designing, and reviewing marketing strategies in real time (Clark, 2001; Maddox, 2001). These extranets are more complex than publishing extranets since they must enable two-way interaction with outside users in addition to document sharing.

The most advanced extranets are transactional in nature and facilitate electronic commerce. These extranets are designed to support online transaction processing (for example, sales) and typically provide EDI services through the public Internet. Transactional extranets are complex to design and implement and require more extensive security.

Inasmuch as the design of an extranet includes both an analysis of the business processes and the technical logistics, individuals who possess an understanding of the business as well as technical experts will need to work together to define the extranet applications, who will be invited to participate from the outside, and how the extranet will be deployed. Failure of extranet projects is often attributed to insufficient involvement of both business and technical people working together on the extranet project design (Covill, 1998).

Determining the Return on Investment

Early estimates from marketing surveys have indicated that the return on investment (ROI) for extranets can be high, especially if the extranet is integrated into business processes where the focus is on cost reduction (Pfaffenberger, 1998). However, managers are often challenged to present cost justifications for new capital projects, especially in the areas of information technology. An extranet application is likely to be no exception.

Return on investment is one method to justify investments in information technology and to determine the priorities for project implementation. The ROI formula compares the benefits of a potential project with its related costs. The formula seems very simple; however, determining the benefits and costs of new technology deployment is not so straightforward. The cost of new hardware and software will be obvious, but estimating the indirect costs and potential savings of an extranet will likely be more of an art than a science.

Bort and Felix (1997) defined types of extranet costs in several broad areas, including equipment, security, development, training, management, support, and taxes. While the extranet might utilize existing networking hardware, such as internal LANs or the Internet gateway, major hardware expenses will likely be required in the areas of additional servers, firewalls, and routers. Defining the purpose for an extranet will inform decision making regarding the hardware needed for the project. For example, a low capacity publishing-type extranet will require less expensive server and firewall protection than a transactional extranet that may require virtual private networking (VPN) security for transmission of sensitive data.

Software for the extranet site must also be considered. While hypertext markup language and more recently extensible markup language might be useful for small-scale extranets, larger and more complex extranets will likely require an investment in content management software.

Consideration must also be given to personnel needs in the deployment of an extranet. Additional personnel may be required to design the extranet application, or to manage the content of the extranet site. Even if existing personnel will be assigned responsibility for management of the extranet, some cost will be incurred for the time current employees will not be available for other projects.

On the benefits side, extranets typically do not generate revenue, but rather they provide organizations with the opportunity to reduce costs. It should be noted, however, that if an extranet is part of an organization's e-commerce strategy, then some portion of the revenue generation should be attributed to the deployment of the extranet (Bort & Felix, 1997).

Some of the potential cost savings for the organization can be estimated in terms of reduced paper costs and improved process efficiencies (Pfaffenberger, 1998). Additional savings could be expected and should be estimated in the following areas where extranets can reduce:

customer service costs by providing customers with direct access to information resources and thereby reduce the demand for customer service personnel;

phone, fax, mail, and shipping costs;

time and cost for product procurement;

training costs; and

any private network charges by using the public Internet and VPN technologies rather than private solutions for EDI.

Additional benefits from extranets, such as customer relationship management and access to information, are even more difficult to measure. An extranet for customers is one way to develop a tighter linkage and potentially gain a competitive advantage. Enhanced communication and access to information in real time are additional benefits within the organization that can result in internal efficiencies. Moreover, organizations may see additional benefits in terms of increased customer loyalty, improved image, and even customer recognition for improved service (Ling & Yen, 2001).

Content Management Strategy

Once the purpose for an extranet is established, developing and managing content on the extranet become priorities. Unlike intranets where content is often developed in a distributed fashion for internal use, content on an extranet is shared with strategic partners and customers outside of the organization. Because of this outside access, the content must be presented in a useable form, and links to resources must be operable. If information on an extranet site is incomplete, outdated, or missing, the organization's credibility and possibly sales will be lost (Reimers, 2001).

The importance of content management cannot be overemphasized. Failures of extranet sites can often be traced to insufficient content management and extranet site development. Successful extranets ensure that their sites offer valuable content and functionality so the users can find what they are looking for quickly. In addition, the extranet site must be fast and reliable.

Selecting products and services for developing and managing the content of an extranet site is a real challenge because of the evolving state of technology. Software tools will be required for Web authoring and content management of the extranet site, database integration tools, auditing tools, and performance evaluation tools. Forrester

Research, in their rankings of content management vendors in March 2001, suggested three questions that should be asked in selecting content management software (Wilkoff, 2001). First, the dominant content source should be identified since this source will require high integration with the content management application. Second, content managers should be determined. Identifying the right content management software will depend on whether the content is going to be managed by a centralized information technology department or delegated to nontechnical business users. Finally, the audience of the content should be identified. If the target audience consists of consumers, then they will require greater performance than typical extranet users. If the target audience of the extranet is buyers, on the other hand, then they will likely require personalization of content. (For listings of vendors and products, see Bort & Felix, 1997; Loshin, 1997; Szuprowicz, 1998.)

In developing extranet applications, consideration should be given to how much of the intranet content will be mirrored on the extranet site. Duplication of extensive databases and applications on an extranet can create problems with synchronizing the data stored on the intranet behind the firewall versus the data provided on the extranet. For this reason, organizations may consider providing trusted partners with direct access to selected intranet servers, although this is more often the exception than the rule because of obvious security concerns.

Security of the Extranet

Security is a critical aspect of extranet development, which extends to both the organization and its partners (Phaltankar, 2000). Security issues must be considered through the design, implementation, and management of any extranet applications (Loshin, 1997). Developing a security plan for an extranet application should begin with a risk assessment to identify the potential sources of threat to the network, how likely these threats are to occur, and the investment (cost) in security that will be required. The level of security investment will vary depending on the nature of the extranet application, the threats of intrusion, and the sensitivity of the information shared on the extranet. Extranet security should consider authentication and access control, privacy, and data integrity (Meyer, 2000).

Access Control

Access control provides or denies access to the network and is usually implemented through deployment of a firewall. A firewall alone, however, is not a sufficient security strategy. Providing confidentiality of information while it is in transit over the public Internet can occur through encryption strategies. User authentication can take place in part at the firewall, but it is usually handled by the application service. User control, however, is almost always handled by the application service (Meyer, 2000).

Authentication

Authentication defines the external population that is permitted to access the extranet and ensures that the external interaction with the extranet is coming from an authorized sourse. Assigning account names and passwords to extranet users is the typical method for implementing authentication. The identity of an extranet user is then confirmed when they present both the account name and password to the host application. Static passwords are not only the simplest way to control access, but also the least secure (Covill, 1998; Szuprowicz, 1998). Moreover, the password if transmitted over the Internet in text (versus encrypted) can be compromised by a password "sniffing" application. A security policy can resolve this potential security breach by using one-time only passwords that expire once a user has been authenticated or encrypting the password before transmission over the Internet (Phaltankar, 2000).

Other forms of authentication are available depending on the level of security required for the extranet. Source address authentication, for example, would authenticate an extranet user based on the IP address of their Internet connection (Phaltankar, 2000; Wack, Cutler, & Pole, 2001). This technique is typically used on intranets to restrict access based on those IP addresses that are internal to an organization. This method of authentication could be sufficient for small-scale extranets with few external organizations where the static IP addresses are known. This technique would not be ideal for global access to an extranet where the IP address of the user is unknown in advance.

Other forms of authentication include tokens, digital signatures, or smartcards to establish the identity of users (Loshin, 1997; Phaltankar, 2000; Pfaffenberger, 1998). Token authentication schemes typically involve two levels of authentication. The first requirement is a personal identification number (PIN), while the second form of authentication is a number displayed on the token card. For example, the SecurID card from Security Dynamics displays a unique number every 60 seconds, which in combination with the user's PIN forms a unique password. The advantage of using tokens or other authentication schemes is that authentication requires two pieces of information: something from the card and something from the user. Without both the card and the valid PIN, authentication will not occur. Of course these additional techniques require additional investments in hardware and software to manage user accounts and for password synchronization between the extranet server and the user's token card (Phaltankar, 2000).

Privacy

Privacy of communication with the extranet and exchange of data are typically implemented through an encryption technique. Encryption is the process of scrambling data before transmission over the public Internet. Several techniques are available, such as the public and private-key cryptography systems developed by RSA (see http://www.rsasecurity.com).

Exchanging information over the public Internet encounters the risk of alteration during transmission. While encryption is an effective strategy to protect the confidentiality of information during transmission, it does not prevent interception and alteration or guarantee that the data packets are received intact.

Data Integrity

Data integrity provides the assurance that the data transmitted over the public Internet are not modified in any way. Data integrity can be implemented by cryptographic checksums and hashing (Phaltankar, 2000). Hash functions create a fixed-size hash value that is a compressed, nonreadable form of a variable-length data message. One-way hash functions are preferred because the hash value cannot be reversed to reconstruct the original message (Covill, 1998). If the same hash value is not generated by the receiver when the data are received, the integrity of the data packet becomes questionable.

For most applications requiring privacy, encryption techniques such as a secure sockets layer (SSL) connection will provide sufficient data privacy and integrity. Virtually all Web browsers recognize the SSL protocol, among others, for data encryption. An organization can provide its business partners with access to an internal application via SSL and password authentication, similar to the way e-commerce Web sites currently interact with consumers (Wilson, 2001).

For sensitive applications or the exchange of strategic data, organizations may require additional security precautions, such as the creation of virtual private networks.

Virtual Private Network

A virtual private network is a technique of implementing secure connections over the public Internet through encryption and authentication schemes. VPNs can also be used to provide secure access to organizational networks for employees who are remote from the organization or perhaps telecommuting. The creation of private networks is also possible through leased lines, X.25 packet-oriented public data networks, or other private IP networks that are available from major service providers. The use of VPNs is becoming increasingly popular because of the cost savings over implementing private networks or remote access facilities, such as modem pools (Pfaffenberger, 1998; Salamone, 2000).

VPNs require firewalls that support VPN technologies. The VPN adds additional protection for the confidentiality and integrity of data transmitted over the Internet by using additional protocols and encryption. The VPN creates a “tunnel” through the Internet that establishes a secured connection from the user site to the extranet. Different protocols are available for implementing a VPN. One popular set of protocols is known as IPSec, the Internet security protocol (Wack, et al., 2001). Other current VPN protocols include point-to-point tunneling protocol, which is a Microsoft standard, and the layer 2 tunneling protocol.

The state of VPN technology and standards for communication protocols are still evolving. The disadvantage of implementing a VPN solution is that it almost always requires the same vendor hardware and software product on both ends of the connection, from the organization hosting the extranet and its business partners. It may not be possible, therefore, to require an organization’s partners to use one particular vendor’s VPN application (Wilson, 2001).

In addition to these security mechanisms, the placement of the extranet within the organizational network can affect security. Early implementations of extranets provided direct access to intranet servers to select outside parties. More recently, a “middle ground” between the public Internet and the internal intranet has become the home of the extranet. This middle ground, called the demilitarized zone (DMZ), is a physical space between the public Internet and the firewall of the intranet (Wack et al., 2001). By placing the extranet in the DMZ, the intranet is protected from potential security breaches by outside parties, both from business partners and potential hackers. Figure 2 depicts an organizational extranet within the DMZ. The DMZ is also used for public access servers, and again restricts the intranet to internal access. One potential disadvantage to this approach is mirroring content from the intranet onto the physically separate extranet that will require synchronization between the servers. The advantage of enhanced security of the intranet, however, would seem to outweigh the disadvantage of synchronization.

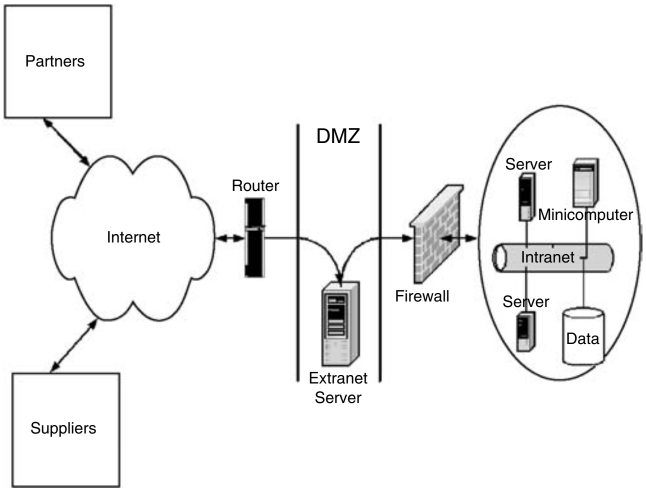


Figure 2: Organizational extranet within the DMZ.

Beyond the security of using the public Internet for business transactions, establishing business partnerships involves careful planning as well. Information technology professionals can design security systems to protect intranet resources from hackers on the public Internet. However, with extranets, the greater security threat may well be from within—disgruntled employees or business partners who have access to internal information (Marcinkowski, 2001). An organization’s extranet partners may come and go, and therefore consideration must be given to guarding against potential misuse of internal information. Moreover, the security of the extranet may be compromised by insufficient security at the partner’s site; that is, a security breach at a partner’s site also exposes the organization’s internal resources made available on the extranet (Schwartz, 2000). Weiler (2001) suggested several steps to safeguarding internal resources:

verifying the security of internal systems before exposing them to extranet partners;

developing contracts between business partners that address expectations, process, hiring background checks, and security administration;

developing human resources notification systems for events such as employee termination;

examining security administration systems; and

planning strategies based on risk management and emerging business requirements rather than on technology.

Outsourcing the Extranet

For organizations that do not have the internal expertise, the time, or the interest to develop and manage an extranet application, extranet service providers may provide a solution (Harreld, 2000). Outsourcing the extranet design, implementation, and management provides the organization with the expertise needed to deploy an extranet while freeing the organization to focus on content management concerns.

Several organizations provide outsourcing resources and technology for deployment of organizational extranets (see, for example, Szuprowicz, 1998). These services typically include extranet development, integration of existing intranets, facilities management, and security. Larger Internet service providers now provide extranet hosting services.

AVOIDING FAILURE IN EXTRANET DEPLOYMENT

Several reasons have been identified for the potential failure of extranet development and implementation (Pfaffenberger, 1998). Because the organization is exposed to external clients, suppliers, or partners through the deployment of an extranet, serious consideration is needed to minimize the potential for a failed extranet.

Potential reasons for extranet failure include:

Selecting technology for the extranet that is not compatible with organizational standards and expertise. If the extranet is going to be managed by internal personnel, then hardware and software decisions need to be mindful of the expertise of existing personnel.

Designing an extranet in the absence of a well-focused plan.

Anticipating potential resistance within the organization and reacting accordingly.

Failing to plan for increased demand and additional capacity. In planning for an extranet, the server architecture should be scalable so additional capacity can be provided as demand for services increases.

Excluding critical stakeholders in the planning process. The success of an extranet will be determined by its usefulness. This includes understanding the needs of business partners who will be accessing the extranet, such as suppliers or customers. A cross-organizational team is ideal for defining the services of an extranet.

Failing to plan and establish content standards and guidelines. Extranets are created for external audiences and therefore makes content guidelines more critical than for internal networks. The organization's image and credibility are at stake if the content provided on the extranet site is not current, reliable, and useful to the external audiences. In a 1999 study of 30 business-to-business sites by Forrester Research, each of the sites failed basic tests of value, ease, and reliability (Sonderegger, 1999). Major problems included missing content, limited functionality, and frequent errors. Planning content for an extranet should be designed with customers, not companies, in mind using a scenario-centered design methodology.

Neglecting to identify who is responsible for the extranet. Extranets are likely to span departmental functions and boundaries and therefore some department will need to be assigned the responsibility for extranet management and coordination of content. In addition, managing the content of an extranet is a time consuming task. As much as 70 to 80% of an extranet's recurring costs are attributable to content management (Pfaffenberger, 1998).

CONCLUSION

Extranets can fundamentally change how organizations share internal resources and interact with outside suppliers, customers, and strategic business partners. The implementation of an extranet can be effective in developing tighter linkages in supply chain management and customer relationship management. The current trends suggest an increasing movement toward developing these linkages through extranet applications. Extranets provide a cost effective means to improve collaboration and decrease the time to market for new products. Moreover, extranets can increase internal efficiencies within the organization. As the extranet examples have shown, applications for extranet deployment are not isolated to one type of industry. Extranets can be effective in seizing a competitive advantage in virtually any type of industry.

The implementation of an extranet solution for any business requires careful planning and a shift in business and management philosophy. Moreover, opening the corporate intranet requires careful assessment of security precautions, defining the external audiences who should have access, and establishing policies for content management.

GLOSSARY

Electronic data interchange A computer-to-computer exchange of business information in a standard electronic format between business partners.

Extranet A collaborative network that uses Internet technology to extend intranet application and services to selected entities external to the organization, such as suppliers, customers, or other businesses that share common goals.

Firewall A device that protects internal networks from unauthorized access from external sources, typically the Internet.

Internet A global network of interconnected networks, providing connection for private and public networks in one cohesive unit.

Intranet A private enterprise network that uses the Internet and Web technologies for information gathering and distribution within the organization.

Virtual private network A technique of implementing secure connections over the public Internet through encryption and authentication schemes.

CROSS REFERENCES

See *Business-to-Business (B2B) Electronic Commerce; Electronic Data Interchange (EDI); Intranets; Supply Chain Management; Virtual Private Networks: Internet Protocol (IP) Based.*

REFERENCES

Anderson, H. (1997). The rise of the extranet. *PC Today*. Retrieved January 7, 2002, from http://www.pctoday.com/editorial/goingonline/970235a.html.

Baker, S. (2000). Getting the most from your intranet and extranet strategies. *The Journal of Business Strategy*, *21* (4), 40–43.

Bort, J., & Felix, B. (1997). *Building an extranet: Connect your intranet with vendors and customers*. New York: Wiley.

Clark, P. (2001, July 16). Extranet pays off in savings for Ketchum. *B to B*, *86* (14), 6.

Covill, J. (1998). *Implementing extranets: The Internet as a virtual private network*. Boston, MA: Digital Press.

Harreld, H. (2000, May 22). Inside extranets. *Federal Computer Week*.

Hills, M. (1998, April 20). Intranets and extranets offer some competitive advantages. *Dallas Business Journal*.

Koch, C. (2000, October 1). Supply chain management: The big payoff. *CIO Magazine*.

Lewis, D. (2001a, January 29). Extranet helps Conoco strengthen retail bonds. *Internetweek, 70*.

Lewis, D. (2001b, May 21). Managed-care firm takes lead with diverse extranet. *Internetweek, 50*.

Ling, R., & Yen, D. (2001). Extranet: A new wave of Internet. *S.A.M. Advanced Management Journal*, *66* (2), 39–44.

Lloyd, P., & Boyle, P. (1998). *Web-weaving: Intranets, extranets, and strategic alliances*. Oxford: Butterworth-Heinemann.

Loshin, P. (1997). *Extranet design and implementation*. San Francisco: Network Press.

Maddox, K. (2001, June 25). Shop turns to collaboration tool. *B to B*, *86* (13), 10.

Marcinkowski, S. (2001). Extranets: The weakest link and security. *SANS Institute.* Retrieved January 8, 2002, from http://rr.sans.org/securitybasics/extranets.php

Messmer, E. (2000, November 13). SunLife extranet to woo insurance brokers. *Network World,* 48–50.

Meyer, K. (2000). Building extranet communities. *Telecommunications*, *34* (6), 77–78.

Moozakis, C. (2001, July 23). Coors looks for a silver bullet—Brewer turns to Web system to track orders and forecast demand with help from distributors. *Internetweek*.

Orlov, L. (1999, August). Surviving the extranet shakeout. *The Forrester Report*.

Pfaffenberger, B. (1998). *Building a strategic extranet*. Foster City, CA: IDG.

Phaltankar, K. (2000). *Practical guide for implementing secure intranets and extranets*. Boston, MA: Artech House.

Reimers, B. (2001, August 13). Content management: Integrate to dominate. *Internetweek,* 29–30.

Salamone, S. (2000, May 8). VPNs enter—The extranet realm—IT managers planning to link business partners can't resist the cost benefits and security VPNs offer. *Internetweek*.

Schwartz, M. (2000, October 2). Good fences, good neighbors. *Computerworld*.

Sonderegger, P. (1999, December). Why most B-to-B sites fail. *The Forrester Report*.

Szuprowicz, B. (1998). *Extranets and intranets: E-commerce business strategies for the future*. Charleston, SC: Computer Technology Research Corp.

Wack, J., Cutler, K., & Pole, J. (2001). *Guidelines on firewalls and firewall policy*. Washington, D.C.: National Institute of Standards and Technology, U.S. Department of Commerce.

Weiler, R. (2001, May 14). Integrating partners carries risk. *InformationWeek, 108*.

Wilkoff, N. (2001, March 1). Content management tech rankings: One size doesn't fit all. *Forrester TechRankings Tech Insight*.

Wilson, T. (2001, May 28). VPNs don't fly outside firewalls. *Internetweek*.

F

Feasibility of Global E-business Projects

Peter Raven, *Seattle University*
C. Patrick Fleenor, *Seattle University*

INTRODUCTION

Businesses realized the commercial value of the Internet shortly after the introduction of the World Wide Web in about 1994 in Switzerland. The graphics capabilities of the Web encouraged many people, even those without technical skills, to utilize the Internet. As a result, the Internet has grown more rapidly than any other commercialized media technology, including radio and television. As the number of people accessing the Internet has grown, the Internet has become a global phenomenon. But why should an e-business consider expanding globally? In a phrase, "that's where the money is!"

A Web site oriented to North Americans reaches only 4% of the world's population and 20% of the global economy. The U.S. share of the global Internet population will decrease from 36% in 2001 to about 24% in 2005 (Dunlap, 2001). Half of all online revenue will come from outside the United States by 2004. Clearly, e-business opportunities will increasingly be global. Unfortunately, two-thirds of U.S. firms are not prepared to do business in the global online market space.

There are many compelling reasons to consider global e-business, not the least of which is the very rapid rate of growth in online populations overseas, some of which have doubled in the past year (Dunlap, 2001). Global markets are often used to hedge economic conditions. For example, while the U.S. may be in an economic slump, other countries may be growing, allowing companies to maintain or even increase sales. Often, the first company into a new foreign market may realize a competitive advantage. Beating competitors to lucrative overseas markets has far-reaching implications in developing new markets and being able to sustain current markets. Progressive firms are increasingly interested in globalizing, but the nuances of conducting global e-business often escape them. Doing global business on the Internet provides many opportunities for product differentiation, creating new forms of complex transactions, simplifying routine purchases, and many other possibilities (Leamer & Storper, 2001). This chapter explores the feasibility of global e-business projects and focuses on global strategies for commercially using the Internet. To answer the title's implied question, is it feasible to develop global e-business? The answer is a resounding *yes*, but it is not always easy or straightforward.

BACKGROUND

In this section we examine the growth of e-business and e-business models to help build a framework for developing global strategies. Although the terms e-business and e-commerce are often used interchangeably, we define *e-business* as all the activities a firm can use on the Internet to facilitate business and *e-commerce* as strictly transactions via a Web site. E-business activities include information dissemination and collection, advertising, promotion, communication, transactions (e-commerce), and many other processes.

Growth of E-business

E-business has grown with the Web. With more than half a billion people online (Table 1), many businesses think it imperative to incorporate the Internet into their strategies. Once a firm is on the Internet, it also has a global presence, but merely having a global presence does not imply that it is ready for global e-business!

Planning for an effective global e-business requires companies to incorporate global Internet strategies. There is considerable anecdotal evidence to illustrate the problems of not developing a good Internet strategy. Neglecting to plan for foreign visitors is one of the more obvious. Evidence from Web site logs suggests that about a third of all visitors to U.S. e-business sites are from overseas (Gutzman, 2000). Ignoring these foreign visitors results in lost business. Further, some Internet firms have even been given a solid order from overseas customers, but have lacked the knowledge to fulfill it, resulting in lost

Table 1 Global Online Population ("Educated guess" as of May 2002)

	February 2002	May 2002
World Total	544.2 million	580.78 million
Africa	4.15	6.31
Asia/Pacific	157.49	167.86
Europe	171.35	185.83
Middle East	4.65	5.12
Canada & USA	181.23	182.67
Latin America	25.33	32.99

Source: NUA, 2002a.

sales and perhaps ill will. Good business practices are especially critical for success in global markets.

Internet Business Models

Several business models exist on the Internet, but four dominate: Business-to-consumer (B2C), business-to-business (B2B), consumer-to-consumer (C2C), and business-to-government or government-to-business.

Business-to-consumer models have the most public visibility and are what most people consider as business on the Internet. B2C models are essentially online storefronts, but there are many variations, ranging from strictly informational sites all the way to highly interactive and personalized transactional sites. Although statistics of Internet usage are suspect due to different measures, the numbers provide some insight. In the U.S., 84% of Internet users have purchased something online (NUA, 2002b). Even though B2C e-commerce is only about 1% of all retailing, 49% of Internet users bought online in 2001 and spending online has increased threefold since 1999, probably due to increased consumer confidence and perceived convenience. Interestingly, consumers express considerably more trust in some institutions than others. A national survey of Internet users revealed that 68% of them trusted information from small businesses almost always or most of the time, but only 32% held that much confidence in large corporations (Princeton Survey Research Associates, 2002).

Business-to-business e-commerce dwarfs that of B2C. B2B models range from procurement (including electronic data interchange) to supply chain management (SCM) to customer relationship management (CRM). B2B models include both informational and transactional activities between businesses. International Data Corporation (IDC) forecasts global B2B revenues to climb from (in U.S. dollars) $282 billion in 2000 to $4.3 trillion in 2005, a compound annual growth rate (CAGR) of 73% (IDC, 2002). The potential for growth is evident by estimates that online e-commerce was only about 2% of all U.S. B2B trade in 2001 and that only 11% of U.S. companies have fully implemented e-business strategies (Cyberatlas, 2002).

While most B2B transactions occur in the U.S., B2B e-businesses in Western Europe and the Asia–Pacific region are expected to grow rapidly. For example, B2B e-commerce was about $9.2 billion in Asia in 1999, but it is expected be $1 trillion in 2004 (Lewis, 2000). IDC predicts the B2B market in the Asia–Pacific region (excluding Japan) to grow at an exceptional rate over the next few years to about $500 billion by 2005, as online access improves and firms become increasingly interconnected (IDC, 2001). Cyberatlas (2002) predicts a CAGR of 68% from 2001 to 2005 in the U.S., Western Europe at 91%, and the Asia–Pacific region at 109%. As can be seen, B2B e-business is expected to be the focus of most of the rapid growth of global e-business in the near future.

Consumer-to-consumer e-business is most evident in the U.S. in auction sites such as eBay and barter sites such as SwapVillage. The size of this market is difficult to measure, since some goods are exchanged, rather than purchased; thus figures reported underestimate the true value of this activity. However, eBayis considered the largest C2C site with 2001 sales of $748.8 million and a growth rate of almost 74% in 2001 (Hoover's Online, 2002). This model has increasingly been adopted in some European countries where auctions are already common and require few new consumer behaviors.

Business-to-government or *government-to-business* is also a growing trend. Many governments encourage the use of the Internet for both internal government communications and between businesses and consumers. The United States took an early lead in this direction during the Clinton Administration with that administration's framework for global e-commerce making government agencies accessible on the Internet (Clinton & Gore, 1997). As an example, the U.S. Internal Revenue Service encourages the filing of income taxes electronically. Taxes may be paid by credit card and refunds deposited directly into a bank account. Finland and Singapore are other countries where a number of government services are available online, with some transactions permitted.

Companies select strategies for their Internet presence—from "pure play," where a company only has an Internet presence, to a "bricks and clicks" model, where a company has both an online and an offline presence. Firms may mix a bricks and clicks model in one market and a pure play in other global markets, depending on whether they already have a physical presence in the market and on their strategy for the market. Our discussion will focus primarily on online B2C and B2B models, with either pure play or bricks and clicks strategies assumed, as these are most important internationally. While there are obvious differences between B2C and B2B, there are also many similarities and many companies employ both models. Manufacturing firms, for example, may use B2B in their SCM and B2C in e-tailing products to final consumers. Therefore, we consider B2B and B2C jointly.

Factors Influencing Global E-business Strategy

Despite the great promise of the Internet for equalization across countries and markets, e-business has not yet fulfilled that promise. Indicators identifying new markets for e-commerce are not well established, but some have gained greater acceptance than others. However, these indicators may not provide the specificity needed to accurately identify potential markets for individual firms. For example, the U.S. has had a tremendous growth in e-commerce due to a homogeneous language and good infrastructure, and because it is a large country (Krück & Papenbrock, 2000). Although most regions of the world do not benefit from all of these important factors, other issues may suggest e-business potential.

Not all countries are equally well connected to the Internet and firms utilize the Internet to different degrees within countries and between countries. Industry Canada, for example, proposed a model adopted by the United Nations Conference on Trade and Development as a starting point for assessing e-business readiness of a country (UNCTAD, 2001). Included in the model are

Readiness of its people, businesses, economy, and infrastructure to undertake e-commerce—of primary interest for countries in early stages of e-business activity.

Intensity with which information and communication technologies are used within a country—of interest for countries where e-business is becoming more prevalent.

Impact of e-business on national economies and business activities—of interest for countries where e-business is well developed.

This model provides an initial, broad screening device that firms can use to understand the current level of e-business of a country before examining the potential more closely.

Many factors influence Internet usage and global e-commerce, just as they do any international trade initiative. Some factors are common to both offline and online market development, but those specific to the Internet include personal computer (PC) penetration, telecommunications infrastructure, and political, economic, and cultural issues related to the Internet and specific to each country. For example, Rasmusson (2000) suggests that wealth, education levels, telecom infrastructure, and PC penetration influence Internet usage in Europe. Ebusinessforum.com (2001b) publishes a ranking of countries based on their "e-readiness." This site is enlightening in that it weights countries on six categories.

Connectivity (30%): adequacy of telecommunications and Internet infrastructure, including fixed and mobile services, affordability, and availability.

Business environment (20%): expected attractiveness of the business environment in the next five years.

E-commerce consumer and business adoption (20%): including payment and logistics systems.

Legal and regulatory environment (15%): including a legal framework involving e-business, support for legal transactions, and digital signatures.

Supporting e-services (10%): intermediaries, portals, Web hosting firms, application service providers, and others.

Social and cultural infrastructure (5%): education, literacy, entrepreneurship, risk-taking, and proclivity for business innovation.

That study concluded that active government support is often needed to assure affordable Internet access. In this regard, large countries do not necessarily have an advantage over smaller countries. For example, India and China both have large populations and, while Internet usage is growing rapidly, these countries suffer from poverty, illiteracy, and poor information and communications technology (ICT) infrastructure. On the other hand, wealthy countries do not always have an advantage over less wealthy countries. More important to e-business readiness, it seems, is a country's embracing of high tech industries and broadband connectivity.

Many of the measures used by E-businessforum.com and others are obvious, but it is necessary to discuss in greater detail PC penetration, infrastructure, and regulatory issues.

PC Penetration

The degree to which personal computers have been adopted is frequently used as a key indicator of e-commerce readiness. PC penetration is highly correlated with income level (Rabe, 2001). For e-commerce to succeed, some threshold level of income is necessary, not only to purchase a PC and products online, but also to access the Internet. PCs are the dominant means of connection and likely to be so for some time. GartnerG2 predicts that in Europe, even with its high penetration of interactive TV and wireless connections, 73% of online shopping will still be done on PCs by 2005 (Regan, 2002b). As might be expected, PC penetration varies widely around the world. In Asia, for example, it ranges from 17/1000 in Thailand to 217/1000 in Singapore (Polster & Trinh, 2000). In the U.S. PC penetration is about 60% of households, with more than 70% in some U.S. cities (Scarborough Research, 2000). However, PC penetration is not a complete measure of Internet access, as office and home computers could be shared, Internet cafes may be popular, and wireless devices can be used.

Infrastructure

Internet transactions require a modern telecommunications infrastructure. Some countries have sophisticated telecommunications systems, which facilitate Internet connectivity. Others, however, have neglected or have been unable to build the necessary systems. Data describing the level of Internet use in a country is often questioned because the estimates vary considerably and may have been published for political or other purposes. Several indicators of telecommunications infrastructure exist to help companies realistically evaluate connectivity. When taken together, they provide a fair indication of current and potential Internet use in particular: telephone penetration, cost of connection, number of Internet service providers (ISPs), and level of broadband usage. Telephone penetration is a reasonable indicator of Internet connectivity because much of the world connects to the Internet through regular telephone lines. Related to telephone penetration is the cost of being connected through telephone lines (WTM, 2001a; Rabe, 2001). Except for Finland and Denmark, European countries pay much higher connection fees than in the U.S., where the average is $25 per month for unlimited access. ISP charges in France average $45 a month and in Germany $68 for just 20 hours. To reduce connection costs, some European ISPs offer free services, taking a portion of the telecommunications fees. These prices are due partly to telephone monopolies and price gouging (Singh, Jayashankar, & Singh, 2001). In addition, many telephone systems charge a toll per unit time of usage. The combination of connection charges and use charges tends to inhibit the usage of the Internet in many countries and, by extension, reduces e-commerce activity. Crispin (2000) indicates that government policies can also severely hinder e-business. In Thailand for example, all long distance telephone and Internet connections pass through the Communications Authority of Thailand (CAT), a monopoly resulting in limited choices, high fees, and historically poor service.

The number of ISPs provides an indication of the extent of commercial Internet use in a country. Presumably, the greater the number of ISPs, the more people within a country utilize the Internet.

Broadband facilitates Internet connectivity, but broadband infrastructure varies a great deal globally. Singapore, Hong Kong, and South Korea are broadly wired for broadband (McKinsey, 2001). South Korea is now the world's leading broadband market with 40% of households connected to the Internet, half by broadband (Dodgson, 2001). Broadband is not yet very common in Europe, but it is growing in popularity.

Economic, Regulatory, and Cultural Influences

The society within which businesses operate influences how business functions and performs. Some social aspects relevant to e-business include the following.

Economic influences are probably the most generalizable because they impact all businesses to some extent. A recession affects both offline and online businesses. The feasibility of e-business is necessarily influenced by the economic conditions in a target country. Economic trends, infrastructure development, income levels, and many other factors influence a country's potential for e-business. For example, the recent economic downturn in Asia has affected all businesses.

Regulation of the Internet is generally not well developed, with most countries either leaving the Internet alone or just contemplating regulations. The European Union (EU) has been the most active in regulating Internet activities, primarily in privacy issues. In May 2001, the European Commission prepared an eEurope action plan with three main objectives: to provide inexpensive, fast, and secure online access; to invest in IT and information society skills; and to promote the use of the Internet (Ebusinessforum.com, 2001a). Recently, a Norwegian was imprisoned for posting racist and anti-Semitic propaganda on a Web site (Seattle Post-Intelligencer, 2000).

The integrity of the institutional environment, especially the "rule of law," is important for the development of e-business (Oxley & Yeung, 2001). Businesses and consumers need to be confident that a firm will deliver satisfactory performance and that there is legal recourse if it does not. While physical infrastructure is most important for explaining *variations* in Internet use, the Oxley and Yeung (2001) data support that e-commerce *activity* depends primarily on the rule of law, and secondarily on the availability of payment methods, such as credit cards. In countries without a strong rule of law tradition, personal relationships become more important and customers are often leery of dealing with faceless strangers over the Internet. In such societies, close-knit trading communities developed to strengthen discipline through informal sanctions. E-commerce is not likely to flourish in its present state in such an environment.

Regulating access to the Internet occurs in several countries. Some countries make a concerted effort to protect their citizens and/or political regimes from accessing the full Internet. China has been very concerned about isolating the country from the rest of the network. Chinese users are encouraged to use the Internet, but access to overseas sites is strictly controlled and the information users post online is closely monitored (Economist, 2001). Singapore, Saudi Arabia, Syria, and the United Arab Emirates filter and censor Internet content. South Korea bans gambling sites. In Iran it is illegal for children to use the Internet and access providers must filter any immoral and anti-Iranian material. In our view, attempts at broad control will ultimately fail. Several countries still attempt to control access to television broadcasts, but the advent of satellite TV has made full control nearly impossible. Internet access will undoubtedly follow a similar development.

Culture and sensitivity to cultural differences play critical roles in successful international business. Understanding how the Web fits into a country's culture is necessary to forming successful customer relations (Rasmusson, 2000). Sites targeting foreign consumers should appear friendly to them, which often requires translating the site into the local language. Eighty percent of Web content is now delivered in English, but almost half of current users do not read English or they prefer to use their own languages. As more non-English speakers go online, translating sites into local languages becomes increasingly important for assuring customer satisfaction. Different languages and cultural values compound the complexity of e-business overseas. An even bigger barrier may be the attitude of business and government entities (Singh, Jayashankar, & Singh, 2001). For example, many European CEOs understand the power of the Internet, but fail to fully exploit it. They see the Internet more as a tool for improving existing business models, rather than a revolutionary device requiring a completely different mindset. In the end, they may be proved right, but this attitude perhaps partially accounts for the uneven acceptance and use of the Internet in industrial Europe.

In addition to language and attitudes, other cultural behaviors such as risk aversion and lifestyle differences may impact success on the Web (Singh et al., 2001). The use of Hofstede's (1980) cultural dimensions can help e-businesses better understand their customers. For example, cultures rating high on uncertainty avoidance and low on individualism are likely to be difficult markets for introducing consumer e-business.

In a recent empirical study, Lynch and Beck (2001) profile Internet buyers in 20 countries in major regions of the world. Their study compares two views of global Internet customers: (1) that there is essentially a world culture of Internet users allowing for standardization of Internet offerings and (2) that there are cultural differences between local markets, suggesting that standardization is infrequently promising (i.e., the standardize/customize dilemma). Their findings indicate that North American Internet users were unique among respondents in that they had a greater affinity for the Internet, less fear of it, enjoyed shopping more on it, and typically looked for branded products. The authors concluded that building a standardized Web site based on North American preferences would not be advisable because of profound cultural differences. Asians, for example, like to shop and look for brands, but have the least favorable attitudes toward the Internet and low intentions to purchase on it. While these differences were noted, the authors also found few differences between countries in a particular region (North America, Europe, and Asia/Pacific).

Global Internet Law

There are no political or natural boundaries applicable to the Internet; so a crime or action committed in cyberspace is not clearly limited to one country or region of the world. In addition, the open infrastructure of the Internet has led to a number of laws in various jurisdictions. The communication and exchange of information cannot be centralized or monopolized on the Internet, as is possible in traditional communication systems. For example, less than a dozen U.S. corporations own and operate 90% of the mass media controlling almost all of America's newspapers, magazines, TV and radio stations, books, records, movies, videos, wire services, and photo agencies (Bagdikian, 1997). The Internet is much more fragmented and loosely controlled by many organizations, including businesses, governmental organizations, individuals, nongovernment organizations, associations, and many others. Nevertheless, countries and international organizations are developing legal frameworks to ensure an efficient and fair use of the available information.

The Internet and its cross-border activities challenge the traditional state-based precepts of private international law. The range of actual and potential disputes affects many aspects in the private and public arenas. Specifically regarding international business, these disputes include e-commerce, taxation, freedom of expression and other human rights, jurisdiction, privacy and data security, and intellectual property rights. The necessity for providing a framework for reliable Internet legislation, adjudication, and enforcement is crucial for global trading. States need to agree on identification of new crimes, jurisdictional rights, and criminal liabilities regarding communications on the Internet. For a basis, states may use legal models of the law of the sea and/or outer space regulations, which deal with similar issues.

Federal governments can regulate a large portion of the Internet, but international e-commerce relies more on international agreements and treaties for legal basis. A variety of mechanisms exist in the international legal systems that provide continuing frameworks for preparatory works in treaty negotiation. The Hague Conference on Private International Law, the International Telecommunication Union, the World Trade Organization, the World Intellectual Property Organization, and other United Nations organizations have permanent secretariats which maintain evolving agendas referring to possibilities for negotiation of new treaties. For example, The Hague Conference on Private International Law attempted in 1999 to establish common grounds for jurisdiction and applicable law arising out of electronic commerce and Internet transactions (The Hague Conference, 1999).

Global Internet law deals with three different issues created by the Internet: first, the Internet facilitates the negotiations of treaty-based systems; second, it alters the balance of interests that shapes the political dynamics determining the content of international law; and third, the Internet's global character challenges traditional state-based precepts of private international law. One of the main reasons is the low economic barrier that allows for access and participation of more people in the discussion of policymaking. Consequently, weaker groups in the political arena can easily mobilize people across national borders. The Internet drastically reduces search costs, thereby making it possible for groups advocating for international legislation to communicate, organize, and bring political pressure to bear. Nongovernmental organizations can easily direct their activities and coordinate through e-mail and the Web. In addition, the Internet aids in detecting violations and mobilizes sanctions on the violators (Perritt, 2000).

The issues are less problematic for e-commerce because the governing law is usually the respective commercial contracts law of a nation. B2B has been fairly easy to deal with, but B2C and C2C have created enormous legal challenges. Private individuals constitute a free-rider problem because they are interested in acquiring information for little or no cost and they cannot effectively be prevented from doing so. Two of the most obvious examples are Napster.com and Morpheus.com, which allowed Web users to download free copies of copyright-protected works. On the other hand, businesses, including e-businesses, have contractual obligations toward each other as there is usually an exchange of some value between the parties. The contract issues are different for private individuals—they do not usually enter into legal contracts with businesses with specified terms and obligations online.

The United States and the European Union represent two different approaches for regulating the Internet. The EU has enacted tough privacy legislation for data collected through the Internet. The EU prohibits data on EU citizens from leaving the EU to countries that do not have privacy standards for Internet data equal to those of the EU—the emphasis here is on individual privacy. In contrast, U.S. privacy policy is not as comprehensive, as the U.S. targets specific areas only when considered needed. U.S. Internet privacy policy is market-oriented. Firms can buy and sell data collected from cookies and mouse clicks to other firms and regulate their conduct themselves. In order to bridge these different privacy approaches and facilitate U.S. organizations' compliance with the EU directive, the U.S. Department of Commerce, in consultation with the European Commission, developed a "safe harbor" framework (U.S. Department of Commerce, 2002). The Safe Harbor framework sets out seven principles in order to be eligible for doing business in the EU. These principles include

Notice: providing notice to individuals that data are being collected and how the information will be used.

Choice: individuals must be given the opportunity to "opt out" and choose not to have their personal information disclosed to a third party.

Onward transfer: to disclose information to a third party, organizations must subscribe to the choice and notice principles.

Access: individuals are to have access to their personal information and the ability to amend incorrect information.

Security: organizations must take reasonable precautions with personal information.

Data integrity: personal information must be relevant and reliable.

Enforcement: recourse and other mechanisms for complaints and verification must be in place.

Firms complying with Safe Harbor are theoretically eligible to transmit data from the EU to their U.S. offices, as well as to receive other legal benefits.

Since the terrorist attacks of September 11, 2001, privacy issues have changed somewhat. The EU has considered making cyber attacks punishable as a terrorist offense. Germany has reduced restraints on government interception of communications. Canada and Australia have introduced laws to redefine terrorist activity and grant powers of surveillance to security agencies if terrorist activity is suspected. Singapore introduced a model data protection code in February 2002 and many other countries have initiated specific anti-terrorism measures. In addition, the worldwide economic slowdown and terrorism anxiety have increased the demand for Web conferencing technologies. As business people travel less, communication is increasingly through electronic means (Geralds, 2002). This, of course, has important implications for global e-businesses. Firms that have geared up for Internet communications and transactions with their global subsidiaries, branches, suppliers, and customers should be at a competitive advantage over their rivals who cling to traditional communication and transaction methods.

Regional Growth of E-business

In this section we examine regional variations in the growth of e-business and examine the critical e-business issues in each region. Our examination of regions is not exhaustive, but, rather, illustrative of regional variations in e-business. By 2005, nearly one billion people (15% of the world population) will be using the Internet (Ecommerce, 2001a). With more than 100 million new users on the Web each year, the Internet will become widespread and not dominated by a single region. However, there are numerous regional differences, many of which will likely persist. We begin with the birthplace of the Internet, the United States and North America.

North America

Growth of Internet use in the U.S. continues, but at a somewhat slower rate than in the past few years. PC penetration and Internet connectivity are in the mature phase of their life cycles, as indicated by their use by large percentages of the population and decreasing rates of Internet growth. In 2000, 54 million U.S. households (51%) had one or more computers and 80% of households (44 million) with computers had at least one member using the Internet at home (Newburger, 2001).

The future for Internet usage is encouraging, though, as 90% of school-age children have access to a computer, either at home or in school, and weekly use has increased (Newburger, 2001). In addition, children increasingly turn on their computers, rather than radios or televisions, to get news and information. Women, minorities, retirees, the less affluent, and the less educated have all increased their usage of the Internet in the past few years (Pastore, 2001).

The Internet is increasingly an important part of the B2B market in the U.S. A recent study found that 45% of large U.S. companies spending $100 million or more per year on supplies used the Internet in the fourth quarter of 2001. This was up sharply from the previous quarter at 28% (Regan, 2002a). Most U.S. corporate executives (53%) see the Internet as either "very important" or "critical" and 87% indicated it was "important" to their purchasing and cost-savings plans in 2002.

Western Europe

The number of Internet users in Western Europe has now overtaken that of the United States (Upton, 2002). Western Europe contributes about 29.8% of global Internet users, while the U.S. has 29.2%. However, Western Europe lags the U.S. in e-commerce revenue—25.7% of the total versus 43.7%. Internet use has grown rapidly in Europe, but somewhat differently from the U.S. Differences in telecommunications charges, landline availability, and other factors caused Western Europe to lag behind North America in Internet usage, especially e-commerce (Greenberg and Spiegel, 2000). Further, there is a reluctance to buy online because of perceived risk of credit card fraud and a mistrust of online shops. Western Europe is characterized by different cultures and languages that have inhibited uniform growth in technological advances. There is also a disparity in Internet connectivity. Connectivity is greatest in the north and progressively diminishes as one travels further south. As a result, the Nordic countries are considerably ahead of other European countries, particularly southern European countries, and even the U.S., in e-business.

Nevertheless, in their e-Europe Initiative of 1999 EU governments stress the need for every European citizen to be online as soon as possible (Liff, 2000). Importantly, the EU has recognized the potential of the Internet to be socially inclusive, building consumer trust and social cohesion (eEurope, 2002). As the e-Europe Initiative progresses through different stages of development, both governments and consumers will be encouraged to utilize the Internet for e-business and other activities. The advent of the euro as common currency in most of the EU will remove even more barriers to e-business (Grant, 2002).

Nordic countries lead Europe in Internet connectivity and Internet usage there is growing, as is online purchasing. Scandinavian countries are the most wired nations in the world. They are also relatively homogeneous markets, making them ideal for a number of marketing initiatives, including call and customer service centers. Digital signatures are legally acceptable, indicating a readiness for e-contracts. While there are individual variations, foreign firms doing business in these countries are expected to pay value added taxes, unless the level of business is below some threshold, when it becomes the obligation of the consumer.

Finland is, of course, the home of Nokia, a world-class marketer of wireless phones and appliances, and the Finnish government encourages the use of Internet commercial activities, including the issuing of national e-identification cards to facilitate online commerce. In a country of about 5.2 million people, 42% of households had PCs in 1999 and 2.2 million were online in 2000 (Ebusinessforum.com, 2002). About 70% of the population in 2000 had cell phones and the number of cell phones has surpassed the number of fixed telephone lines. By 2003 it is estimated that more mobile phones than PCs will be connected to the Internet.

The infrastructure in Finland is in place for e-business, but Finns seem reluctant to shop online. Jupiter Communications expected only 36.8% of the population to purchase online in 2002. Most businesses with more than 20 employees had Internet access in 1999 and about 52% of Finnish firms ordered goods and services online. Thus, this country is poised for e-business growth and online volume has increased considerably with monthly turnover of B2C e-commerce estimated at (in Finnish Markkaa) 113 million in February 2000 (Ebusinessforum.com, 2002). B2B e-commerce is estimated to be at least 10 times that amount.

Mobile commerce (m-commerce) is attracting much interest and its potential is very large due to the expected rapid growth in mobile phone use in the next few years. In a recent move to capitalize on the global growth of mobile phones and the huge potential for m-commerce, Nokia will license its source code controlling such functions as mobile Web browsing and the popular short messaging software to a consortium of mobile phone manufacturers and operators (Brewin, 2001). The mobile software components will initially include multimedia messaging, digital rights management for music and video, subscriber authentication, and XHTML for Web browsing. The authentication software is seen as being able to jump-start m-commerce.

Sweden is the largest of the Scandinavian markets and has taken the lead in IT technology evolution (Yorgey, 2001). Relatively low telecommunications costs, a high GNP/capita, and favorable social policies drive this technology evolution. In this country of almost 9 million, 71% of the population is online and 60% of the population uses cell phones (Ebusinessforum.com, 2002). B2B e-commerce is growing, especially in the music and steel industries. Sweden is a mature Internet market with most of the purchasing options found in the U.S. B2B, B2C, and C2C are typical business models, but the Swedes are also launching consumer driven and alternative pricing sites, including “group shopping” sites where vendors drop the price after a critical mass of buyers request it.

Norway has considerable wealth generated by its offshore petroleum and gas resources. Norway is well wired and e-commerce has taken off with little hindering it (Ebusinessforum.com, 2002). With 500,000 installed lines, Norway has the highest rate of integrated services digital network penetration in world. E-commerce spending is expected to reach $4.7 billion in 2002 because online consumers can buy everything from food to books to electricity on the Internet. As is typical in Nordic countries, online banking is available through even the smaller banks as well as Internet-only banks. In the B2B area, Norwegian businesses readily use the Internet for procurement, logistics, and data transmission. The government is planning to allow firms to pay taxes, fees, and customs duties electronically. The oil industry utilizes e-commerce extensively in online auctioning for procurement, providing management information, and integrated systems through third-party service providers (Kingsley, 2000). M-commerce is, of course, taking off rapidly.

With the largest population of any country in Europe, Germany is an important market, but e-commerce represents a relatively small part of its economy. However, with 1.4 million Internet hosts in 1998, 25% of the population owning a PC, and 10% able to access the Internet, the Internet future appears bright. E-commerce spending is expected to reach $62.8 billion in 2002, about twice that of 2001 (Ebusinessforum.com, 2002). The German government continues to liberalize the Internet and telecommunications sectors and to implement EU directives (Lubben & Karenfort, 2001). Many of the regulations are designed to enhance consumer confidence in e-commerce. For example, the Distance Selling Act applies to all contracts between sellers and consumers by distance communication, including the Internet. It gives consumers the right to withdraw from such a contract for a period of time if the seller does not provide sufficient details about the product/service and consumers’ legal rights. However, Germany has strict regulations regarding offline and online store hours.

Although France has had online capabilities for about two decades through the state-run Minitel system, their Internet capabilities have not kept pace. The Minitel service is quite expensive and does not connect easily to the Internet, so it is of limited e-business value. Only 12% of the French population is connected to the Internet, but these people are avid users, primarily of chat programs. One of the factors inhibiting commercial use of the Internet is the cost of connection, which can run from $2.50–$4.00 per hour, not including ISP charges. Europeans in general, and the French in particular, are somewhat reluctant to purchase online because of fear of credit card fraud (Greenberg & Spiegel, 2000).

The U.K. is more similar to the U.S. than with its European neighbors in its use of ICT for e-business. Regulations in the U.K. and Sweden tend to be lighter than in other European countries, facilitating Internet shopping (Ecommerce, 2001b). The British have taken to the Internet, both consumers and businesses, and are much more likely to use credit cards for payment than other Europeans.

Southern European countries, such as Spain, Portugal, and Italy, lag behind their northern neighbors, but there are encouraging signs even here. These countries have seen some of the highest rates of average growth in Internet usage in Europe and the IDC believes that they will reach levels comparable to the more advanced European countries in coming years (Skipper-Pedersen, 2001).

European Internet penetration seems to have plateaued at about 38% in December 2001 (Ebusiness-forum, 2002). The slower uptake of ICT by consumers is partly a result of persistent high costs of computer connections, even though dialup costs are decreasing since the telecommunications markets have been opening up. Broadband costs are significant, running €45–€60 per month. B2C e-commerce is developing more slowly than predicted, leaving some doubt if the EU will meet its objective of making it "the world's most dynamic knowledge based economy" by 2010. In 2002, Germany is expected to dominate e-commerce sales with $20 billion, followed by the UK at $9 billion, and France with less than $4 billion (Regan, 2002b). However, European online purchasing habits differ between countries. Swedes, the Swiss, Spaniards, and Italians prefer to buy with cash, the French with checks, and Germans, Swiss, and Austrians want an invoice option to pay. Only the British include credit cards among their top three payment methods.

Asia–Pacific Region

Asians seem to adopt technology long before adapting it to e-business. One study found that small- and medium-sized enterprises (SMEs) face a number of barriers in adopting e-commerce in Asia. One of the most important barriers is the low use of e-commerce by consumers and companies within Asia–Pacific Economic Cooperation (APEC) countries, particularly the less developed member nations. This inhibits SMEs from selling online to consumers and conducting e-commerce with other companies (PWC, 2000).

The Boston Consulting Group estimates that Asian B2C revenues more than doubled in 2000 and will grow by more than 100% per year, led by the banking and finance industry (Peng, 2002). There is an average of only 0.01 PCs/person in Asia, compared to 0.40 in the U.S. As a result, average online retail spending is quite low, but there is wide variance. For example, on average, Indonesians annually spend $0.01/person online, Indians $0.03, Thais $0.20, and Malays $0.50. Even though the affluent societies of Taiwan, Singapore, and Hong Kong are very well connected to the Internet, people there would rather shop in malls, where they can touch and compare products, as shopping is considered a pastime. Asians are also wary of using credit cards online. To reduce this concern, VISA now offers a "Verified by VISA Program" using personalized passwords to serve as additional verification.

Despite these issues, there is substantial opportunity for e-business in Asia, especially in China and South Korea. Although business portals have difficulty generating revenue, e-commerce revenue is expected to grow to more than $16 billion over the next few years (China Daily, 2002). Firms have found creative ways of overcoming the lack of credit card penetration. For example, Jiangsu Province is testing a new "cyber-notary" system to reduce the risks of identity theft in Internet bidding (Ji, 2002). Dell Computers is allowing customers to pay on delivery or through banks (Einhorn, 2001). Dell sees China as a country with 1.3 billion people and less than 10% PC penetration, but closer to 30% penetration in wealthier cities. Dell's strategy is to reduce the price of computers to the point where 85% of the Chinese urban population can afford one.

South Korea makes a concerted effort to increase connectivity and now it has the highest broadband penetration in the world, with more than half of South Korea's 15 million households having broadband and 60% of the population with cell phones. The Seoul suburb of Seongnam plans to be the world's first digital city (Ihlwan, 2002). Broadband connections there are expected to do away with cash and credit cards. Citizens will be able to pay for purchases at every store via digital cell phones.

Japan is noted as the Asian leader in technological innovations and for its high penetration of NTT DoCoMo's i-mode, a cell phone messaging system, but less so for its Internet connectivity. Despite the high levels of mobile phone penetration, most Japanese connect to the Internet via PCs (75%). Internet penetration remains under 20% in Japan and will likely remain below the level of other countries in the region, such as Singapore (41% penetration), Australia (32%), Taiwan (29%), Hong Kong (22.6%), and South Korea (21.2%) (Cheung, 2001c). However, Internet penetration in Japan is projected to grow to 29% by 2004, with e-commerce revenues growing proportionately. B2B e-business has the potential to modify traditional Japanese business practices, especially the inefficient and expensive distribution system through disintermediation. Credit cards are not widely used in Japan, but innovative ways around that are sometimes used. For example, the local convenience store sometimes acts as Internet kiosk, cash payment collector, and pickup point for products ordered online. Japan's share of e-commerce revenues for 2000 were 70% of Asia–Pacific's total e-commerce market and will likely fall to 60% in 2004, as the rest of Asia catches up. Japan's B2C e-commerce revenues are expected to climb to $24 billion and B2B to more than $180 billion by the end of 2004. A survey conducted jointly by Accenture, the Japan Ministry of Economy Trade and Industry (METI), and the Electronic Commerce Promotion Council of Japan (ECOM), is even more optimistic, projecting Japan's B2B market to grow to $955 billion and the B2C market to $115 billion by the end of 2005 (Cheung, 2001b).

Much of the projected growth in Japan and Asia will be fueled by m-commerce because the Asian cellular market is one of the most advanced in the world. In many Asian countries, poor telecommunications infrastructure and expensive telephone connectivity have driven the move to wireless. With the use of 3G services in Japan and elsewhere, m-commerce will likely grow rapidly. Japan and South Korea will probably take the lead in this area (Cheung, 2001a). The ARC group predicted that Japan alone will account for 25% of world wireless subscribers by 2003, and Asia (including Japan) for more than 39%.

Other Areas

Many other areas of the world are not yet ready for e-business or are just getting started. Due to insufficient infrastructure to support widespread B2C e-business and a small market because of low discretionary income, e-business in these areas will undoubtedly start as B2B.

However, it is premature to dismiss B2C entirely, as its potential is growing.

In Latin America, for example, e-commerce revenues are expected to grow to $67 billion by 2005, mostly in B2B e-commerce (Ecommerce, 2001c). Brazil dominates Latin America with almost 4 million Internet users, or 40% of this region's total. PC penetration is low—from 3% in Peru to 10% in Argentina. Because of fixed line infrastructure limitations, Latin America is increasingly turning to wireless connections. By 2005, more than 50 million Internet users in Latin America will access the Net via mobile devices.

Africa and the Middle East have low PC and Internet penetration levels. However, even in these countries the use of the Internet is growing, primarily in the B2B area. Unfortunately, the struggle to reach sustainability has eluded many e-businesses. In South Africa, for example, the 2.4 million users are probably insufficient to support many e-businesses (Ebusinessforum.com, 2002). In both Egypt and Saudi Arabia, Internet penetration is very low and costly and many people use Internet cafes to access the Internet. Internet cafes allow information exchange, but tend to inhibit e-commerce per se, as transactions are less convenient compared to using individually owned PCs. While there is potential for growth in these countries, it will be slow. The use of mobile phones, however, has grown much faster.

Eastern European countries, especially Hungary, the Czech Republic, and Poland are much brighter spots in the growing Internet market space. Generally higher levels of education, acceptance of technology, and growing economies in these countries suggest that businesses and consumers in these countries will be receptive to e-business. A potential impediment in these countries, as elsewhere, is concern about data security. A national sample of American Internet users revealed that 65% worry a lot or somewhat that someone might obtain their credit card number and misuse the information (Princeton Survey Research Associates, 2002). While many companies are attracted to the Eastern European markets because of proximity to the EU and lower prices, these markets are still evolving and the B2B market is underdeveloped (Drummondi, 2001).

Global E-business Strategies

Merely having a Web site does not constitute a global strategy. Substantial effort is required in planning and executing a successful global e-business strategy. We now focus on advice for prospective e-entrepreneurs.

Business Models

Some firms, especially those with a bricks and clicks model, use their Web site primarily as an informational brochure (*brochureware*), describing the firm and its products and services. They expect interested customers to go to a physical store or contact the firm via e-mail or telephone to request more information. This has some appeal as an international strategy because it is relatively low risk. There are drawbacks, though, because a visitor to the site can easily click somewhere else. Beyond brochureware, an advanced step includes interactivity, personalization, and data collection about visitors. This increases the "stickiness" (length of stay) of a site, increases return visitors, and provides valuable information about a visitor's online behavior. A further stage is transactional, where customers order and pay for the product/service online. This stage involves two-way communication between the customer and the company. Any of these stages may be appropriate for a particular international market, but the most appropriate Web site will depend on the market, the product/service, and the goals and objectives of the firm. In global e-business, the challenges are in communicating with the right customers, focusing on their needs, providing efficient fulfillment, and ensuring secure payments.

Setting Goals and Objectives

Almost anyone can put up a static Web site for little cost, but designing an appropriate e-business site requires setting goals and objectives for the site and skill in its execution. The investment in developing an interactive, sophisticated, transactional site is usually very high, as are the risks of doing the job poorly. In addition, the countries on which a firm is focused, as well as the target segment within those countries, must be clearly identified before a site is launched. Misidentifying the target market may well misposition the Web site and the product, and thus reduce revenue.

Goals and objectives help determine the type of site, whether simple brochureware or more complex interactive and transactional sites. For example, an informational site designed to complement offline stores should describe products, the firm, retail locations, and contact information. Measuring the success of such a site can be as simple as counting the number of e-mails or phone calls generated in a period of time. However, an objective of increasing sales through a Web site requires an appropriately sophisticated transactional site. Metrics can determine number of pages downloaded, click-throughs, sales per unique visitor, or other appropriate measures. Data may be collected anytime a visitor clicks on a Web site, but many firms do not complete the process by analyzing the data and utilizing that information in their online strategies. The analysis of these metrics is crucial to successful e-commerce. Learning which pages are downloaded, which ads are seen, site stickiness, etc., provides invaluable information to aid e-businesses in developing appropriate online strategies for their target markets.

Consumer Behavior, Relations, and Culture

Understanding a customer's culture and their online behaviors are crucial for effective global e-business. Country culture and consumer behavior are related. Web site design must respect the target culture. This may be as simple as providing a language option or as complex as avoiding culturally sensitive issues. For example, certain products are offensive to some cultures, such as pork or alcohol in Islamic cultures. Colors mean different things to other cultures, such as white representing death and mourning to Muslims and Chinese. The number "four" symbolizes funerals to Japanese and Chinese, so items packaged in fours or the display of four objects should be modified for sites targeting the Japanese. Cultural nuances and symbols are difficult to understand and predict for nonnatives, so one should have a sample of the target audience and culture preview and screen the site for culturally sensitive items and correct language translation before launch.

Consumer behavior online differs in some respects from more traditional purchasing activity. Theorists believe many consumers go through a mental process in purchasing products, especially expensive products or those with a high social cost (e.g., image). This decision-making process occurs both offline and online, but is instructive in developing appropriate Web sites. The purchase decision process consists of searching for ways to solve a perceived problem (e.g., how should I resolve my need to get to work on time?), analyzing the choices (e.g., public transportation, buying a car), determining prospective solutions (e.g., Honda, Toyota), and then making the purchase (e.g., based on price, performance, service, perceived value). Focusing on each of the steps in the purchase process can facilitate online purchasing. If international consumers are in the information-gathering stage, a firm's site should focus on providing information to help make the purchase decision. Different cultures may focus on different data in making the decision, which complicates this process, but makes research very important. For example, purchasing an automobile may involve an information search on the engineering and technical aspects of the vehicle in one culture, but the options or image of the vehicle in another. The importance of providing the right information is that your competitors are just a mouse click away.

B2B consumer behavior is less variable than in B2C transactions. Businesses typically purchase on well-defined specifications, easily disseminated via a Web site or e-mail. However, people in different cultures require different levels of human contact. People in high context countries (e.g., Asians, Latin Americans), for example, may require personal phone calls, e-mail, or even personal visits in order to maintain the desired relationships.

Adapting to Cultures

A major value of the Internet lies in its community building. A company with an Internet presence and global strategy makes the commitment to participate in this increasingly global community. Still, assuming that the whole online world is now part of a company's market is unrealistic; thus it is important to specifically identify those markets the firm wishes to, and is able to, service well. It is important to establish partners in target countries in order to serve the market more specifically. In this way, adapting the site to a market's needs becomes easier. For example, target country partners will be able to identify culturally sensitive issues to avoid on a Web site, language nuances that may not be obvious, and specific needs of target customers.

Global Customer Relationship Management

CRM refers to the visible front-end operations that create and maintain customer satisfaction. Offline, it includes such activities as keeping in touch with customers through customer service representatives, maintaining store relationships, and so forth. Online, CRM requires presenting a consistent image and following up inquiries with e-mail and customer satisfaction inquiries. CRM can be integrated with back-end operations such as supply chain management to help the entire supply chain focus on customer needs and expectations creating strong synergy (Strauss & Frost, 2001).

CRM requires understanding the lifetime value of a customer, because it is much less expensive in the long term to retain a good customer than to acquire a new one. As a result, firms invest in keeping good, profitable customers and "firing" less profitable customers. Globally, this translates into managing the costs of acquiring new customers and retaining current customers. In their well-known study, Peppers and Rogers (1996) found that it costs about five times as much to acquire new customers as to retain current customers. The ratio internationally is probably much higher. Consequently, it is critically important to maintain good customer relationships with global customers.

The Global Internet Marketing Mix

The impact of a global Internet strategy on the marketing mix is significant. One of the key advantages for global Internet marketers is the ability to provide mass customization, which refers to the ability to electronically customize some or all of the marketing mix to suit a target market. This allows firms to present a customized offering much less expensively than with traditional methods.

Product Strategies

Products and services acceptable to domestic Internet markets are generally appropriate for global Internet markets. Some consumer products require adaptation for specific markets, but others can be marketed as is. B2B e-marketing can be very efficient in that nonstandardized products can use the Internet to expedite the customization process. Customization is facilitated by allowing design specifications and product modifications to be changed in real time, with approval likely to be much faster than that of traditional methods. Likewise, online auctions and trade exchange sites, such as the Online Technology Exchange and the Chemical Industry Data Exchange, facilitate the important processes of procurement by providing information on offerings and by bringing providers and suitable customers together.

Electronic products are especially well suited for global Internet marketing because the Internet audience already has a higher proportion of technology savvy users and the products require little or no customizing. Consequently, online marketing of electronic hardware has become one of the mainstays of e-commerce. Software is easily marketed on the Web because of the high involvement of users and the opportunity for users to sample the product online, purchase it online, and even have it delivered online.

Distribution and Fulfillment Strategies. The discussion of distribution and fulfillment in e-business is essentially one of electrons vs. atoms, as distribution functions include not only physical distribution of a physical product, but also the locale where the product is offered for sale (Web site or physical store), transfer of information, payments, breaking bulk, assembly, warehousing, etc.

Fulfillment refers to how an order is completed and is directly related to distribution. Distribution and fulfillment strategies are significantly more complex in supplying international markets, requiring well-organized strategies. Distribution encompasses the value chain from manufacturer/producer to the final consumer, but the entire value chain, including the supply chain, should be considered as a whole. Due to the complexity of overseas shipping, especially documentation, many e-commerce firms use third party logistics providers. Firms such as Fed Ex, Airborne Express, and UPS can provide a complete logistical solution for overseas deliveries, but costs must be able to be supported by customers, while the firm maintains profit margins. Price escalation due to high delivery costs may reduce the global marketability of a product on one hand, but on the other may be insignificant compared to the basic cost of a product or the unavailability of substitutes.

One of the promises of the Internet has been disintermediation, in which intermediaries are eliminated from the distribution system, reducing the length of the distribution channel (Leamer & Storper, 2001). The Internet allows customers to buy direct from manufacturers or to eliminate one or more intermediaries. This potentially reduces costs to the consumer, but the functions provided by intermediaries must still be performed, sometimes by the customer. For products or services not requiring physical distribution, such as consultation or software downloads, disintermediation may result in lower costs to the firm and lower prices to consumers. However, for firms with both an online and offline (physical or bricks and mortar) presence, there are good strategic reasons for maintaining similar retail prices online and offline, such as reducing consumer confusion and discontent, and maintaining customer loyalty. For example, if prices are cheaper on the Web site than in the physical retail store, business may be cannibalized from the retail store. Customers may also feel confused and even cheated if there are large discrepancies in price between physical and virtual stores.

Disintermediation sometimes leads to reintermediation. Reintermediation occurs when distribution channels, which have been disintermediated, are reformed in new and unique ways to provide value added services. For example, reintermediated services include auto-locating services, information and e-commerce services in real estate and home services that combine multivendor information, comparison shopping, and sometimes ties to local retailers (Hanson, 2000). Reintermediation strategies could provide much needed innovative services in global markets.

Fulfillment issues are often glossed over in favor of developing glitzy, interactive Web sites. In a bricks and mortar retail outlet, this function is unseen and largely taken for granted by consumers. Stores already have the product in stock, or will order it, and delivery is a behind the scenes issue. For an online retailer, however, fulfillment issues are critical and force companies to examine their entire value chain (Sowinski, 2001). Better integration between supply and demand chains helps to resolve exceptions such as backorders, returns, and incorrect orders, while improving delivery times and customer satisfaction. Fulfillment issues are thus critical for several reasons: first, they lead to consumer satisfaction or dissatisfaction; second, they influence a firm's profitability; and third, they affect returns.

Customer satisfaction is directly related to fulfillment. Of course the product/service must meet customer expectations, but it must also be delivered when and where the customer wants it and in a satisfactory condition. Integrating customer service functions is crucial to online success; poor customer service is detrimental in a global retail marketplace where competition is only a mouse click away (Chow, 2000). E-commerce customers expect speed and quality of delivery. Meeting, and exceeding, customer expectations results in customer satisfaction, repeat business, loyal customers, and profit (Saenz, 2001). However, B2B and B2C firms must solve different fulfillment problems.

B2B customers typically order few items, but in high volume. B2C customers order a variety of items in low volume, such that the frequency of B2C smaller orders can be astronomical and overwhelm even well designed fulfillment systems (Saenz, 2001). The basic fulfillment functions of picking, packaging, and returns are common to both B2B and B2C, but B2C systems must accommodate large numbers of small orders. Efficient fulfillment operations return large dividends in profit and in customer satisfaction, but many e-businesses are ill equipped to perform this function well themselves and, instead, rely on third party providers.

Outsourcing fulfillment makes sense for many e-businesses, but at what point should an e-business consider this option? Aichlmayr (2000) recommended that e-businesses with fewer than 1000 orders per day build their own fulfillment infrastructure. Firms with more than 10,000 orders/day will likely want to exercise greater control by keeping the fulfillment function in-house. E-businesses handling between 1000 and 10,000 orders/day are often better off outsourcing fulfillment with firms such as PFSweb Fulfillment Services, Omni Fulfillment, Connextions Fulfillment Center, i2 Technologies, and many others.

There are several types of third party providers for firms to consider (Aichlmayr, 2000): "Physical Infrastructure Providers" perform warehousing, delivery, and returns functions; "Technology Providers" supply technological infrastructure to trading partners, including order management and virtual supply chain; and "Integrators" provide both technological and physical infrastructure, such as shopping cart technology, credit card processing, tax and freight calculations, and order tracking. Global e-businesses may desire distribution facilities in their major global markets to assure optimum fulfillment and customer satisfaction.

Unfortunately, not all customers are satisfied, and they return their purchases. Traditional retailers (bricks and mortar stores) expect about 10% of purchases to be returned, but e-tailers get about 30% returns on average (Saenz, 2001). For an e-business with a large volume of small orders, returns can seriously affect the bottom line. Global e-businesses can be impacted to an even greater extent than domestic businesses. Part of the problem appears to be communication. Forrester Research indicates that 85% of e-businesses have no automated, real-time connection with any of their business applications.

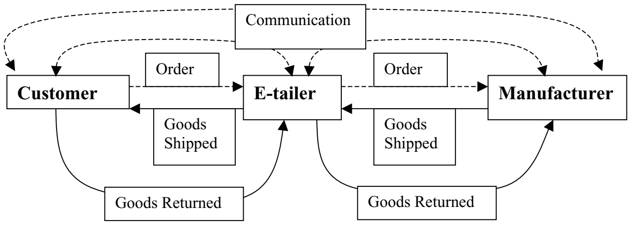


Figure 1: E-commerce fulfillment model.

Further, a Jupiter Communications study showed 42% of Web sites never responded to customer inquiries, took more than five days to do so, or did not offer e-mail responses to problems (Chow, 2000). Clearly, the level of customer service has not kept pace with the growth of e-business. Handling customer inquiries and returns is a critical activity for global e-businesses. Outsourcing this function has been effective for many firms, including Pharmacia & Upjohn, IBM, Hewlett-Packard, Dell, Nokia, Lancome, Tyco International, Sony On-Line Entertainment, and others.

A simplistic model of e-commerce fulfillment is shown in Figure 1 to conceptually illustrate the process. Dashed lines indicate the flow of communications and solid lines the flow of goods/services. The model becomes significantly more complex as the number of intermediaries increases and as customers, retailers, and manufacturers are separated by political and geographic boundaries. For example, while ordering and communications are quite straightforward and very rapid over the Internet, the shipping of physical goods and even some services will be much less so. E-businesses need to consider customs, duties, and logistics as goods are shipped to customers. They also need to manage a returns policy that will meet customer needs without incurring excessive costs. As indicated earlier, outsourcing may facilitate all of these functions.

The increase of B2C e-commerce domestically and internationally will challenge fulfillment operations. One area of fulfillment often overlooked, but capable of having long-term effects on profitability, especially internationally, is packaging. Jupiter Communications reports that 27% of e-tailers spend more than 10% of their revenues on fulfillment, more than half related to packaging (Hogan, 2001). Fewer than 62% of Internet orders are shipped using best practices, and problems include high costs and high rates of return due to broken or damaged products. Consumers are also concerned about the size of package, wanting packages close in size to the actual product. Designing packaging that is effective in protecting the product while reducing material and handling costs, including returns, is challenging, but obviously important. In addition, packaging must not create a bottleneck to rapid delivery.

Pricing Strategies. Internet pricing strategies take on new complexity in international markets. Price floors must take into consideration various influences to price escalation, including transportation, tariffs, duties, and increased channel length. Price ceilings reflect cultural and economic issues and the ability to pay. But, pricing is also strategic, so prices must also reflect competition, demand, perceived value, and a host of other factors. These issues are present in offline international marketing as well, but what makes Internet marketing unique is the ability of customers to rapidly search the Web for competitive prices.

Web scanning is facilitated by use of shopping agents or bots, which scour the Internet, researching products by price and other parameters determined by the bot manager. This requires Internet marketers to continually scan the market space for competitor price changes. This is a significant issue in the EU, where the euro is now the common currency for most member countries. Price transparency and the ability to rapidly search for competitive prices make pricing strategy considerably more important today than in the past.

The Internet allows for real-time pricing, which permits price changes to occur instantaneously (Arend, 2002), emphasizing the need for tracking competitor prices. For example, a marketer pricing a product in France can rapidly determine if competition is changing prices to compete, permitting the marketer to immediately adjust prices on their Web site. Price changes in the past often took considerable time to implement because price schedules had to be changed, catalogs printed, and various channel members informed. The Internet allows all of this to be done very rapidly.

The Internet also facilitates yield/capacity management, sometimes referred to as “perishable asset revenue management.” These practices are most often used in industries with a great deal of volatility in purchases, such as airlines and hotels. The Internet allows the rapid change of prices to help ensure that airplanes fly at or near capacity and that hotels, resorts, and restaurants are filled on weekends and other off-peak times. In the airline industry, tickets purchased online are often less expensive than those purchased through travel agents and e-tickets are even less. Besides reducing transaction costs, the Internet facilitates yield management technology and information dissemination, allowing lower fares through an Internet purchase.

Communication and Promotional Strategies

Advertising and promotion via the Internet are common in the U.S. and can be very effective. Interactivity makes the Internet an important advertising vehicle, as it provides advertisers opportunities to identify customers, differentiate them, and customize purchase and post-purchase service and offerings (Roberts & Ko, 2001). Customers benefit by having greater and quicker access to companies through e-mail, discussion groups, direct ordering, and links to further information. As with offline advertising, online advertising must be sensitive to cultural, language, and consumer behavior differences. Hall’s “silent languages” provide an example. High context/collectivist cultures (e.g., Asians, Latin Americans) use indirect modes of communication and favor emotional appeals reflecting the needs of the group. On the other hand, low context/individualistic cultures, such as North Americans and Scandinavians, use direct modes of communication and seem to prefer high content messages. Which strategy an e-business should use will, of course, depend on the market on which they focus. Web sites designed for Scandinavians would probably not be effective for Latin Americans.

Probably the best-known approach to understanding cultural values in the business literature is the theory of national culture proposed by Geert Hofstede. Hofstede identified four main value dimensions affecting cultures, power distance, uncertainty avoidance, individualism, and masculinity, later adding a fifth, long-term orientation, which better explained Chinese cultural values (Hofstede, 1980, 1991, 1994). National culture is considered the most important influence on international business and Hofstede's work over the years has demonstrated the usefulness of the five value dimensions in a number of cross-national studies that can be generalized to international e-business. For example, cultures low on individualism and high on power distance (e.g., many Asian and Latin American countries) should respond to Web sites emphasizing collectiveness and hierarchy. People in these countries would not respond in the same way to sites emphasizing individualism and egalitarianism, as would Nordic Europeans and North Americans.

Web advertising is growing at a faster rate outside the U.S., but the U.S. is still the dominant Web advertising market (Halprin, 2000). Online advertising in Europe is different from the U.S. due to lower levels of Internet use and broadband penetration, as well as language, culture, and market issues. Because Internet usage is growing rapidly in Europe, online advertising is expected to grow from $901 million in 2000 to $6.4 billion in 2005. Germany is expected to be the largest online advertising market in Europe, followed by the U.K. and the Scandinavian countries, as a group.

Online advertising is also rapidly increasing in Asia, from about $465 million in 2000 to nearly $4 billion by the end of 2004, or about 12.1% of world total (Cheung, 2000). Still, online advertising remains a relatively unproven and distrusted concept in Asia, primarily because of relatively low PC penetration and a lack of credible, consistent data on Web traffic volume. Click-through rates on banners are low. Wireless advertising, on the other hand, is expected to grow strongly, especially in Japan, China, and South Korea.

Unfortunately, little work has been done on estimating the value of Internet advertising internationally, but assuming most domestic online advertising issues apply to global online advertising, we can make some generalizations. First, the Internet allows for very focused communications through e-mail. This is similar to direct marketing, but much more effective because it can be more focused and is relatively inexpensive. Focus is enhanced by e-mail addresses generated by opt-in requests from users. These are exactly the people who will react to the ads and information. E-mail costs about 1/100th that of regular direct mail in the U.S. (Gartner, 2002) and is undoubtedly even less expensive to overseas customers.

Personalized opt-in e-mail generates about 6–7% click-through, compared to 1% for generic e-mail and direct mail. Besides being less expensive and more effective than direct mail, e-mail marketing campaigns can be measured more quickly and easily than traditional direct mail. E-mail campaigns should be attractive for global marketers with the ability to focus on those customers more likely to purchase. E-mail, of course, should be in the language of the customer, correctly translated from the home language, and should have contact information so customers can respond. Banner ads have been shown effective at improving a brand's image and attitude toward a brand, but less effective in inducing purchasing behavior. This is not surprising, as Internet users are familiar with banner ads and less likely to click-through than they were when banner ads were relatively new. However, in some global markets banner ads may be new and exciting and quite effective at creating awareness of a product category and brand.

CONCLUSION

In our examination of the feasibility for global e-business activities we found not only many similarities between regions, but also major differences. North America, especially the United States, has entered a mature phase of e-business. Many other countries are just developing their e-business infrastructure and technology to compete in an increasingly global and technological world. Europe and the Asia–Pacific regions are growing in their e-business capabilities faster than any others and, therefore, appear to be prime areas for e-business market development. Most other areas of the world are growing in Internet connectivity and are potential e-business markets in the near future, primarily for B2B e-business.

Firms considering global e-business must recognize that e-business in foreign markets is likely to be different than domestic e-business. First and foremost is the language issue. Even if a foreign target audience understands English, it may be important to have its home language available on your site, making users feel welcome and signal that you are truly interested in serving them. Other important cultural factors to consider include symbols with specific cultural meanings. We discussed colors and numbers as being symbolic, but there are many more, as well.

Other cultures also have different ways of doing business. Those impacting e-business are the use of credit cards and other e-payment options. It is very important for both parties that credit information be secure. If customers are reluctant to use credit cards, we must find other ways for convenient payment. Innovative ways of serving global customers are likely to encourage not only more sales, but also stronger customer loyalty.

Fulfillment is critical for satisfying customers and for the bottom line. B2C e-businesses can outsource this function if they are not prepared to handle the volume of business a successful global strategy can attain. In any case, long-term profitability and brand image may well rest on this important process. Providing the quality of product or service demanded by the global market space is no longer sufficient—it must also get to its proper destination in the expected time, in the expected condition, and at the expected price, while still providing a profit to the firm. This is not always an easy task.

As with any global effort, proper and timely communication will diminish problems. The Internet facilitates routine communication, but does not substitute for face-to-face communication desired by many cultures, where personal relationships can last and be profitable for a long time.

The outlook for global e-business is encouraging. It will not look exactly like it has in the U.S., but it is growing and will continue to grow. Wireless connectivity in many parts of Europe, Asia, Latin America, and other regions will change the nature of B2C e-business, and probably B2B business, as well. E-business firms that have well-developed business models and a thorough understanding of global markets will be leaders in the new global e-economy.

ACKNOWLEDGMENT

The authors gratefully acknowledge the Institute for Global e-Business and Innovation at Seattle University for its assistance on this project, especially Research Assistants Claudia Zeibe and Tan Chee Ming.

GLOSSARY

Banner ad Rectangular space on a Web site that is paid for by an advertiser.

Bricks and Clicks Retail stores with both online and offline selling.

Broadband High bandwidth necessary for transmission of multimedia content over the Web.

B2B Business to business—marketing of products to businesses and institutions.

B2C Business to consumer—marketing of products to end consumers.

Brochureware Information on a Web site describing a firm's products and services and the firm itself.

Click-through Determined when a Web user clicks on a banner ad that is hyperlinked to the advertiser's site.

CRM Customer relationship management—process of identifying, attracting, differentiating, and retaining customers using digital processes and integration of customer information.

E-business All electronic activities conducted by a business.

E-commerce Buying and selling online.

ICT Information and communications technology.

ISP Internet service provider—a firm with a network of servers, routers, and modems attached to the Internet. It is used by subscribers to connect to the Internet.

Shopping agents (Bots) programs used to allow consumers to scan the Web to compare prices, features, retail locations, etc., of products.

CROSS REFERENCES

See *Business-to-Business (B2B) Internet Business Models; Business-to-Consumer (B2C) Internet Business Models; Customer Relationship Management on the Web; Electronic Commerce and Electronic Business; Global Issues; Intelligent Agents; International Cyberlaw; Politics*.

REFERENCES

Aichlmayr, M. (2000, November). From data to delivery: Finding fulfillment in e-business. *Transportation and Distribution, 41* (11), S3–S10.

Arend, C. (2002, January). Euro to push B2C ecommerce in Western Europe. *IDC*. Retrieved February 4, 2002, from http://www.idc.com

Bagdikian, B. (1997). *The media monopoly: With a new preface on the Internet and telecommunications cartels*. (5th ed.). Boston: Beacon Press.

Brewin, B. (2001, November 19). Nokia rivals team up to develop 3G service aps. *Computerworld, 35* (47), 7.

Cheung, E. (2000, December 1). Internet advertising in Asia. *E Marketer*. Retrieved April 13, 2002, from http://www.emarketer.com

Cheung, E. (2001a, February 22). Asia mcommerce: Hyper or just hype? *E Marketer*. Retrieved April 13, 2002, from http://www.emarketer.com

Cheung, E. (2001b, March 2). The land of the rising ecommerce. *E Marketer*. Retrieved Aprill 13, 2002, from http://www.emarketer.com

Cheung, E. (2001c, May 8). The Japanese Internet market. *E Marketer*. Retrieved April 13, 2002, from http://www.emarketer.com

China Daily. (2002, March 19). eBay snared by country's World Wide Web. *China Daily*. Retrieved March 19, 2002, from http://www.chinadaily.com.cn.

Chow, E. (2000, December). Outsourcing e-commerce fulfillment. *Warehousing Management, 7* (11), WM12–WM13.

Clinton, W. J., & Gore, A., Jr. (1997). *A Framework for Global Electronic Commerce*. Retrieved March 13, 2002, from http://www.chimes.com.au/exposure/bill.html

Crispin, S. W. (2000, September 21). E-commerce emasculated. *Far Eastern Economic Review*. Retrieved April 26, 2001, from http://www.feer.com/_0009_21/p26region.html

Cyberatlas. (2002, March 6). B2B e-commerce headed for trillions. *Cyberatlas*. Retrieved March 13, 2002, from http://www.cyberatlas.internet.com

Dodgson, C. (2001, June). Why Korea? *Communications International*, B20.

Drummondi, N. (2001, August 31). Eastern promise: Companies are flocking to the east in quest of value for money. *E.Business*, 30.

Dunlap, B. (2001, April 5). Why your company should go global now more than ever. *Global Reach*. Retrieved February 27, 2002, from http://glreach.com/eng/ed/art/rep-eur23.php3

Ebusinessforum.com. (2002). Doing ebusiness in.... *Economist Intelligence Unit*. Retrieved March 20, 2002, from http://www.ebusinessforum.com.

Ebusinessforum.com. (2001a). Doing business in the European Union. *Economist Intelligence Unit*. Retrieved December 5, 2001, from http://www.ebusinessforum.com

Ebusinessforum.com. (2001b, May 8). The Economist Intelligence Unit/Pyramid Research e-readiness rankings. *Economist Intelligence Unit*. Retrieved March 20, 2002, from http://www.ebusinessforum.com

Ecommerce. (2001a, June 13). Net users worldwide taking commerce online. *Internet.com*. Retrieved April 11, 2002, from http://www.ecommerce.internet.com

Ecommerce. (2001b, June 7). European e-tailers face regulatory, cultural barriers. *Internet.com*. Retrieved April 11, 2002, from http://www.ecommerce.internet.com

Ecommerce. (2001c, January 19). Wireless access. E-commerce on rise in Latin America. *Internet.com*. Retrieved April 11, 2002, from http://www.ecommerce.internet.com

Economist. (2001, August 11). Leaders: The Internet's new borders. *Economist, 11*, 9–10.

eEurope. (2002, February 3). eEurope: An information society for all. *Information Society Website*. Retrieved February 18, 2002, from http://europaeu.int/ISPO/I_europe.html

Einhorn, B. (2001, December 3). Dell takes a different tack in China. *BusinessWeek Online*. Retrieved January 20, 2002, from http://www.businessweek.com

Gartner. (2002, March 19). GartnerG2 says e-mail marketing campaigns threaten traditional direct mail promotions. *Gartner Research*. Retrieved March 28, 2002, from http://www4.gartner.com

Geralds, J. (2002, June 9). How the wired world was affected. *Informatics Online*. Retrieved September 6, 2002, from http://www.infomaticsonline.co.uk

Grant, E. (2002, February 13). Study: E-commerce to top $1 trillion in 2002. *E-Commerce Times*. Retrieved March 25, 2002, from http://www.ecommercetimes.com

GreenbergGreenberg, P. A., & Spiegel, R. (2000, January 13). European e-commerce still lags behind U.S. ecommerce times. Retrieved March 25, 2002, from http://www.ecommercetimes.com/perl/story/22–3.htm

Gutzman, A. (2000, May 17). Globalization of e-commerce: Are you ready? *E-commerce-guide.com*. Retrieved April 11, 2002, from http://ecommerce.internet.com

The Hague Conference on Private International Law. (1999, November). *Press release*. Retrieved April 11, 2002 from http://www.hcch.net/e/events/press01e.html

Halprin, D. (2000, December 14). Web advertising grows outside the U.S. *E Marketer*. Retrieved April 12, 2002, from http://www.eMarketer.com

Hanson, W. (2000). *Internet marketing*. Cincinnati: South-Western College Publishing.

Hofstede, G. (1980). *Culture's consequences: International differences in work-related values*. Beverly Hills, CA: Sage Publications.

Hofstede, G. (1991). *Cultures and Organizations: Software of the Mind*. London: McGraw-Hill.

Hofstede, G. (1994). Management scientists are human. *Management Science, 40* (1), 4–13.

Hogan, P. B. (2001, October). Packaging: Fast way to improve customer satisfaction and reduce cost of fulfillment. *Material Handling Management, 56* (11), 59–60.

Hoover's Online. (2002). eBay. Inc. Hoover's online. Retrieved February 4, 2002, from http://cobrands.hoovers.com

IDC. (2001, April 23). IDC predicts a boom, not gloom for Asia's B2B ecommerce markets. *IDC Press Release*. Retrieved February 4, 2002, from http://www.idc.com

IDC. (2002). Major growth forecast for B2B revenues. *IDC Research*. Retrieved February 4, 2002, from http://www.nua.com

Ihlwan, M. (2002, February 4). South Korea: A nation of digital guinea pigs. *BusinessWeek Online*. Retrieved February 27, 2002, from http://www.businessweek.com

Ji, B. (2002, February 5). E-credit system undergoes trial run. *China Daily*. Retrieved March 28, 2002, from http://www.chinadaily.com.cn

Kingsley, M. (2000, December). Oil-fired up. *Australian CPA, 70* (11), 20.

Krück, U., & Papenbrock, K. (2000, May 22). Regional starting positions in global e-competition. *Deutsche Bank Research*.

Leamer, E. E., & Storper, M. (2001, Fourth Quarter). The economic geography of the Internet age. *Journal of International Business Studies, 32* (4), 641–665.

Lewis, S. (2000, April). Asia Embraces B2B E-commerce. *Asian Business, 36* (4), 24.

Liff, S. (2000, September/October). Consumer e-commerce: Potential for social inclusion? *Consumer Policy Review, 10*(5), 162–166.

Lubben, N., & Karenfort, J. S. C. (2001, July). Moving forward with deliberate speed? *Corporate Finance, 200,* R38–R40.

Lynch, P. D., & Beck, J. C. (2001, Fourth Quarter). Profiles of Internet buyers in 20 countries: Evidence for region-specific strategies. *Journal of International Business Studies, 32* (4), 725–748.

McKinsey, K. (2001, February 22). Shoppers lost in cyberspace. *Far Eastern Economic Review*. Retrieved February 17, 2001, from http://www.feer.com/_0102_22/p034innov.html

Newburger, E. (2001, October 10). 9-in-10 school-age children have computer access; Internet use pervasive. Census Bureau Reports. *Public Information Office. U.S. Census Bureau*. Retrieved February 11, 2002, from http://www.census.gov

NUA. (2002a). How many online? *NUA Surveys*. Retrieved September 16, 2002, from http://www.nua.com.

NUA. (2002b). Intermarket group: Half of US users bought online in 2001. *NUA Surveys*. Retrieved February 4, 2002, from http://www.nua.com

Oxley, J. E., & Yeung, B. (2001, Fourth Quarter). E-commerce readiness: Institutional environment and international competitiveness. *Journal of International Business Studies, 32* (4), 705–723.

Pastore, M. (2001, May 30). Internet use continues to pervade U.S. life. *Cyberatlas*. Retrieved February 11, 2002, from http://cyberatlas.Internet.com

Peng, S. L. (2002, March 15). E-retailing in Asia will boom. *Internet.com*. Retrieved April 11, 2002, from http://ecommerce.internet.com

Peppers, D., & Rogers, M. (1996). *The one to one future*. New York: Doubleday.

Perritt, H. H., Jr. (2000). The Internet is changing the public international legal system. *Kentucky Law Journal*. University of Kentucky College of Law, 88 KYLJ 885.

Polster, R., & Trinh, T. (2000, November 7). Emerging Asia: From hardware base to e-commerce space? *Deutsche Bank Research, 9*.

Princeton Survey Research Associates. (2002, April 16). A matter of trust: What users want from Web sites: Results of a national survey of Internet users for consumer Webwatch. Retrieved April 25, 2002, from http://www.consumerwebwatch.org/news/

PWC (2000). SMEs face barriers in adopting e-trade. *PriceWatershouseCoopers Website*. Retrieved August 3, 2001, from http://www.pwcglobal.com

Rabe, J. (2001, August 15). The digital divide—Focus on Asia. *Deutsche Bank Research, 17*.

Rasmusson, E. (2000, February). E-commerce around the world. *Sales and Marketing Management, 152*(2), 94.

Regan, K. (2002a, January 17). Report: B2B e-commerce gaining strength. *E-Commerce Times*. Retrieved February 11, 2002, from http://www.ecommercetimes.com

Regan, K. (2002b, March 19). Report: European e-commerce growing fast, but risks remain. *E-Commerce Times*. Retrieved March 25, 2002, from http://www.ecommercetimes.com

Roberts, M. S., & Ko, H. (2001, Spring). Global interactive advertising: Defining what we mean and using what we have learned. *Journal of Interactive Advertising, 1* (2). Retrieved April 9, 2002, from http://jiad.org.vol1.no2/Roberts

Saenz, N., Jr. (2001, May). Picking the best practices for e-fulfillment. *IIE Solutions, 33*(5), 37–40.

Scarborough Research (2000). Three U.S. cities reach 70 percent home PC penetration according to latest Scarborough study. *VNU MI Inc. News*. Retrieved March 17, 2002, from http://www.vnumis.com

Seattle Post-Intelligencer (2000, April 4). Elsewhere: Norway. *Seattle Post-Intelligencer,* A4.

Singh, T. J., Jayashankar, V., & Singh, J. (2001, March–April). E-commerce in the U.S. and Europe —Is Europe ready to compete? *Business Horizons, 44*(22), 6–16.

Skipper-Pedersen, J. (2001, December). IDC's annual econsumer survey—Ready for 2002? *IDC*. Retrieved February 4, 2002, from http://www.idc.com

Sowinski, L. L. (2001, May). International e-commerce fulfillment. *World Trade, 14*(5), 60–62.

Strauss, J., & Frost, R. (2001). *E-marketing*. Upper Saddle River, NJ: Prentice Hall.

UNCTAD (2001). E-commerce and development report 2001. *United Nations Conference on Trade and Development*. Retrieved March 20, 2002, from http://www.unctad.org/ecommerce

Upton, M. (2002, January 3). Western Europe pulls ahead of United States. *eBusiness Trends. ITWorld.com, 89*. Retrieved March 25, 2002, from http://www.idc.com

U.S. Department of Commerce (2002). *Export portal, safe harbor*. Retrieved August 5, 2002, from http://www.export.gov/safeharbor/sh_overview.html

Ward, H. (2001, May 24). Internet access will challenge democracies. *Computer Weekly,* 16.

WTM (2001, June 14). Beyond multilingualism. *World Trade Magazine. Global Online*. Retrieved August 3, 2001, from http://www.worldtrademag.com/

Yorgey, L. A. (2001, November). Sweden takes the lead in the Scandinavian tech revolution. *Target Marketing, 24*(11), 47.

File Types

Jennifer Lagier, *Hartnell College*

INTRODUCTION

Each time an individual searches the Internet, the inquiry results in the retrieval of information in the form of a file. This file may be displayed, downloaded, or executed and may exist in the form of a Web page, image, document, data, software program, or multimedia, such as digitized movies or sounds. The organization of the file is called its format. The choice of file format depends on the kind of information being transmitted and the way it will be used (Hofstetter, 1998).

File formats and their associated file name extensions provide clues about what the file contains—a software program, information that has been compressed, text, image, sound, video, animation, or data. To view, play back, execute, manipulate, interpret, or alter the contents of the file requires different software tools or specific hardware configurations for different file types (Hofstetter, 1998; Ackerman & Hartman, 2001).

Files may interact to produce a displayable Web page, or they may be stored on a Web server and made accessible from a Web page that contains hyperlinks coded to retrieve the requested file. Sometimes files are transferred to and from Web servers using special software designed for this purpose. File transfer software permits the uploading or downloading of files from the Internet without requiring the user to actually open the file.

Some Web pages work together with special script or program files stored on the Web server. Some examples are JavaScript (JS) files, active server pages (ASP), PHP (personal home page) script files, common gateway interface (CGI) files, and cascading style sheet (CSS) files. By themselves, these files are of little use, as they contain formatting or program instructions, not actual Web content. However, when used in conjunction with hypertext markup language (HTML), these separate files can (a) retrieve data from an online database, then create Web pages on the fly to display search results or store and modify information; (b) interact with other programs to perform a variety of online actions; or (c) determine the display characteristics of one or multiple Web pages (http://www.whatis.com).

Text format files include HTML files and document files that can be downloaded or retrieved and displayed. The most common text file types encountered on the Internet are plain text, rich text, and word processing files (Miller, 1997). These files may be identified by their file name extensions: .txt for plain or ASCII text, .rtf for rich text, and .doc or .wps for word processing files. Some text files may be compressed and saved in this format or as an image to save space or retain formatting characteristics.

Data files usually consist of files created by spreadsheet programs. In addition, some Web sites make use of various database programs to create data files. These files are saved in a nontext format that can be translated and displayed by the software programs that originally created them. Newer versions of spreadsheet programs often contain filters that permit the interchange of files created by multiple software programs (Ackerman & Hartman, 2001).

To reduce the time required to download from the Internet and to conserve Web server space, larger files are frequently compressed to create small file sizes. Compressed files must be extracted before they can be used. Some compressed files have been saved as self-extracting executable files. Others require the use of file compression programs to uncompress the file (Ackerman & Hartman, 2001). Some common file name extensions for compressed files are .zip, .gz, .hqx, .sit, and .tar.

The software program used to create these images frequently determines graphic file formats. The most common formats found on the Internet include GIF and JPG or JPEG. The Web browser can display images in these formats without the use of additional software (Miller, 1997). Graphic images stored in bitmapped (.bmp) or tagged image formats (.tif or .tiff) can be downloaded from the Internet, but to be displayed, they require the use of other software programs (Miller, 1997; Ackerman & Hartman, 2001).

Audio files require the use of a computer equipped with a sound card and speakers, as well as media player software to convert the file into sound. Recent versions of Netscape Navigator (7.0) and Microsoft Internet Explorer (6.00) come with the necessary software components to deal with most common audio files. Older versions of browsers may require helper applications to allow the computer and browser to access the appropriate program or action to make use of the file. Some audio files, such as RealAudio and MP3 formats, require use of additional media players or software programs to convert these files into sound (Ackerman & Hartman, 2001). Common file name extensions for audio files include .wav, .au, .snd, .ra, .ram, .aif, .aiff, .aifc, .mid, .midi, and .mp3.

Common multimedia file formats include MPG or MPEG (Moving Picture Expert Group), MOV or QT (QuickTime), RM or RAM (RealVideo), DCR, DIR, and DXR (Shockwave). Multimedia files permit the simultaneous viewing of digitized video accompanied by sound. As with audio files, multimedia files frequently require player applications for conversion to motion and sound.

Server-side files are files saved in a format that allows them to interact with HTML pages. Some of the more common formats include ASP, CGI, JS, and CSS. Files of these formats are stored on a Web server and function in a way that determines the formatting and display characteristics of Web pages. ASP and CGI scripts interact with information from databases, such as online catalogs, archives, and digital library collections.

Although the Internet provides users with a wealth of information, much of which can be downloaded and saved, it is important to consider copyright restrictions that may limit or prohibit copying or distributing files. Copyright laws of the United States, the Universal Copyright Convention, and the Berne Union protect Web pages and their contents (Ackerman & Hartman, 2001).

TEXT FILES

There are four different text formats most commonly available on the Internet: plain text, rich text, hypertext, and word processor files (Hofstetter, 1998). All browsers can display plain text, or ASCII, files. ASCII (American Standard Code for Information Interchange) is the most commonly used text format for Internet or computer files. Text files contain characters that can be viewed or printed, but they do not include formatting, such as special font faces, or styles, such as bold or italics. Files of this type usually have the file name extension .txt (Ackerman & Hartman, 2001).

Hypertext files contain text that has been surrounded with special HTML codes. The HTML coding defines how text will be displayed on the page (font face, font size, font color) as well as whether or not the text will appear in bold, italics, or underlined. HTML coding also permits text to function as a trigger or link that takes the user to another Web site, Web page, or place within the same Web page. Hypertext links can also launch various actions or activate multimedia files. Hypertext files have either .htm or .html file name extensions (Hofstetter, 1998). The World Wide Web Consortium (W3C) maintains a Web site that provides information about HTML standards and specifications (http://www.w3.org/MarkUp/).

Such word processors as Microsoft Word and WordPerfect create files that contain proprietary formatting. The format codes within a word processor file determine font face, font size, margins, borders, headings, footers, and pagination (Hofstetter, 1998). Word processing files have file names that may end in a variety of extensions, such as .doc (Word document file), .wpd (WordPerfect document file), or .rtf (rich text format). Unlike text and hypertext documents that can be opened and read by Web browser software, a browser cannot display word processor files. These files require helper applications in order to be opened and have their contents displayed. Rich text format files contain text that can be translated by multiple word processing programs (Ackerman & Hartman, 2001).

PostScript is a file format invented by Adobe Systems. These files contain text, but they are usually not in a readable form. They contain commands that a printer or display device interprets—commands relating to formatting, fonts, font size, and images within the file (Ackerman & Hartman, 2001). A PostScript file usually has a file name extension of .ps.

Although PostScript is recognized as the industry standard for printing and imaging, portable document format (PDF) files, also invented by Adobe, are more widely used for Internet document exchange (http://www.whatis.com; Miller, 1997). Adobe Acrobat creates PDF files by converting word processor or printed documents into bitmap images. Although the software to create PDF files is not free, Adobe Acrobat Reader software, used to view PDF files, is available as a free download from the Adobe Web site, (http://www.adobe.com/products/acrobat/readstep2.html).

Text files in a variety of formats are used to provide online documentation, product order forms, registration forms, college catalogs, course handouts, newsletters, research articles, dissertations, publications, brochures, and more. If document layout and print formatting are unimportant, ASCII text files provide a simple method of sharing files on the Internet. Word processor text files are an appropriate choice for those working on collaborative projects, editing and augmenting each other's work. Hypertext files are well suited for online course content, interactive Web-based tutorials, online catalogs, forms, or lists of hypertext links. Portable document files duplicate desktop publishing formats, allowing users to view publications exactly as they were designed to appear in print versions. These files may be accessed by using a free downloadable reader and do not require any specialized application programs to open, print, or view.

NUMERICAL DATA FILES

Spreadsheet programs, such as Microsoft Excel and Lotus 1-2-3, produce files containing data represented in a nontext format. Excel files have extensions of .xls. Lotus 1-2-3 files have extensions of either .wks or .wk1. As with word processor files, spreadsheet files need to be opened by the program used to create them. Newer versions of spreadsheet programs contain translation filters that can open and display data with any of the file extensions listed (Ackerman & Hartman, 2001). In addition, Excel software allows files to be saved in HTML format, creating Web pages suitable for viewing on the World Wide Web. Data files are most often used to share or display statistical information, such as census data, mortgage rates, interest tables, and demographic breakdowns.

COMPRESSED FILES

Files are often compressed to save space on a server and to permit transfer that is more rapid over the Internet. One or more files can be packed into a single file. Once the user has downloaded a compressed file, the file will need to be unzipped or extracted to view, display, or use it. Users may download application software from the Internet to perform this function (Ackerman & Hartman, 2001).

Self-extracting archives are compressed files containing multiple application files or a set of data packed into a single file. After downloading a file to the computer's hard drive, the user must launch the archive file, which then unpacks the files and stores them on the computer. Many software companies now distribute application programs over the Internet using self-extracting archives (Hofstetter, 1998). Examples of self-extracting software programs may be found at CNET's http://download.com (Hofstetter, 2001).

A significant number of the files accessed through anonymous file transfer protocols (FTPs) are stored in a compressed format (Ackerman & Hartman, 2001). The most common compressed or archived files are those with extensions of .zip, .gz, .hqx, .sit, or .tar. Shareware programs that work with compressed files are available from http://www.pkware.com, http://www.stuffit.com/expander/index.html, and http://www.winzip.com.

IMAGE FILES

Graphic image files may be stored in a variety of formats. Most Web browsers can display images stored in GIF (graphics interchange format) or JPEG (Joint Photographic Expert Group) format. Files with images in these formats have file name extensions of either .gif, .jpeg, or .jpg.

Tagged image file format (TIF or TIFF) is used to store and exchange high-quality bitmap images. Some browsers are unable to display TIFF files, necessitating the use of software applications that can either display or convert TIFF files to a format such as GIF or JPG, which can be viewed with any browser.

PNG (portable network graphics) is a new standard currently being developed by W3C. The new PNG format would provide a patent-free image file format that would replace GIF and TIFF format files.

BMP (Windows bitmap) images are composed of a grid broken into small squares called pixels. Each pixel contains color and location information. Because each BMP graphic contains a specific number of pixels, magnification can produce image distortion.

GIF

CompuServe Information Services introduced the GIF image format. GIFs were designed to be a platform-independent format designed for slow modem speed transfers (http://www.wdvl.com). GIFs are 8-bit compressed images that can be displayed without any loss of original graphic information. This process is known as lossless compression. GIF files are limited to a palette of 256 colors and are considered one of the standard image formats found on the Internet. A GIF file is encoded in binary as opposed to ASCII, or text, format. Binary files require the use of a software program to interpret their content. Transparent GIFs allow a background image color to be designated as "see through," allowing the illusion of images that float over a Web page's background color, texture, or graphics. Transparent GIFs may also be used to represent type fonts and blank spaces. GIF files can be in GIF 87a or GIF 89a format. The latter permits creation of an interleaved, or animated, image (http://www.whatis.com).

Animated GIFs

An animated GIF appears on a Web page as a moving image. Animated GIFS are created from a series of unique images saved within a single file in GIF 89a format. The animation displays a sequence of images that appear in a specific order. An animated GIF can display its sequenced images once and then stop, or it can loop endlessly (http://www.whatis.com).

JPEG

JPG, or JPEG, images are compressed 24-bit files that can contain over 16 million colors. JPGs are used to display digitized photographs, full color, and naturalistic grayscale images (http://www.wdvl.com). The JPG format allows a range of compression qualities. The greater the compression ratio, the lower the display quality of the image and the smaller the file size. As with the GIF format, JPG is one of the standard Internet image formats (http://www.whatis.com).

PNG

Currently, the W3C is working on a new graphics format called PNG (portable network graphics). This format will handle images containing up to 48 bits of color information per pixel and will eventually replace the older GIF format. As with GIF, PNG image files are lossless, yet will permit compression rates 10%–30% more compressed than the GIF format.

An Internet committee designed the PNG format to provide a patent-free alternative to the GIF. Unlike the GIF format, PNG files not only support transparency, they also allow a range of transparency values. PNG also supports interlacing, gamma correction (the ability to tune the brightness of an image to correspond with specific monitor displays), and the ability to save in true-color, browser safe, or gray-scale formats (http://www.whatis.com).

Mapped Images

Mapped images contain an invisible layer of hyperlinks over the image. By moving the cursor over the image and clicking on various spots within the image, the user triggers the link (Weinman, 1999). Links may retrieve external Web sites, Web pages, places within the same or another Web page, images, or other types of files, such as digitized videos or sound files.

AUDIO FILES

In order to hear sound files from the Internet, the user needs to have a computer equipped with a sound card, speakers, or headphones, as well as a plug-in or helper application that works in conjunction with the Web browser software. When the Web browser encounters a sound file with a WAV, MIDI, NeXT/Sun, MP3, or audio interchange format, it downloads the entire file onto the hard drive of the computer. Depending on the file size and speed of the Internet connection, this may take several minutes. Once the download is completed, the helper application is launched. The Windows Media Player is an example of a helper application that might be used to play back audio files.

The Web browser then loads the sound file into the application. The application plays the sound. A plug-in performs the same function as a helper application only it does so inside the browser window rather than by starting up another software application (Burns, 1999).

Rather than downloading an entire audio file before activating playback, streaming audio downloads a portion of the sound file into a buffer, then continues downloading as playback proceeds. Information is transmitted from the Internet to the user's workstation in a steady, continuous stream. Because many users do not have high-speed Internet connections, streaming technologies are becoming an increasingly popular way to deliver online multimedia (Ackermann & Hartman, 2001; Hofstetter, 2001).

Sound files transmitted over the Internet often have one of the following formats:

Waveform audio format (WAV), a standard format for computers using Microsoft Windows (file names end with .wav).

NeXT/Sun format (file names end with .au or .snd).

RealAudio format (file names end with .ra or .ram).

Audio interchange format (file names end with .aif, .aiff, or .aifc).

Musical instrument digital interface (file names end with .mid or .midi).

MP3 format (a compression system for music files that does not diminish sound quality).

WAV

Microsoft created the WAV audio file format, which has become accepted as the standard PC audio file format. Both PC and Macintosh computers can use WAV files. The WAV file contains raw audio in an uncompressed format, as well as information about the file's number of tracks (mono or stereo), sample rate, and bit depth (http://www.whatis.com).

A waveform audio digitizer can be used to record sound. The waveform describes a unique sound's frequency, amplitude, and harmonic content. A waveform audio digitizer captures sound by sampling a waveform thousands of times per second, then storing the samples within a file format suitable for Internet transmission (Hofstetter, 2001).

AU/SND

NeXT and Sun computers produce sound files in a unique format. The audio file names usually end with the extensions .au or .snd. Audio files with extensions of .snd do not have file headers to indicate different sampling rates and compression formats, thus allowing variable recording and playback rates. Audio files with names ending in .au extensions are recorded and played back at a standard rate. Audio files produced by NeXT and Sun workstations provide a larger dynamic range than normal 8-bit samples, approximately equivalent to 12-bit samples (Hofstetter, 1998).

RA/RAM

RealAudio refers to continuous or streaming sound technology developed by Progressive Networks' RealAudio. RealAudio sound files are stored in a special format optimized for real-time transmission over the Internet and can be recognized by their file name extensions of either .ra or .ram. This is the format used for Internet radio broadcasts. Unlike other sound files, which must be completely downloaded before playback can occur, RealAudio files are downloaded and played simultaneously. RealAudio files must be played through a RealAudio player. The RealOnePlayer is able to read the file stream as it is coming in and begin playing it long before the rest of the file arrives, thus providing continuous, or streaming, audio (Hofstetter, 1998). RealOnePlayer is available as a free download from http://www.real.com.

AIF/AIFF/AIFC

Audio interchange file format (AIF, AIFF, or AIFC) is the format used to create audio files on Macintosh computers. Sound files with the extension of .aifc indicate a Macintosh-produced audio that has been compressed. When used on a PC, these files usually have the file name extension .aif (Hofstetter, 1998).

An AIFF file contains the raw audio data, channel information (monophonic or stereophonic), bit depth, and sample rate. AIFF files do not support data compression, so files in this format are usually large. Files with the extension .aifc indicate a compressed version of the AIFF format. The AIFC format supports compression ratios as high as 6:1 (http://www.whatis.com; Hofstetter, 1998).

MID/MIDI

The musical instrument digital interface (.mid or .midi) audio file format was designed to enable the recording and playing back of music on digital synthesizers supported by many makers of personal computer sound cards (http://www.whatis.com). Unlike other audio formats, MIDI does not directly record the actual sound. Instead, MIDI records information about how the audio file is produced. The information includes note-ons, note-offs, key velocity, pitch bend, and other methods of controlling a synthesizer (Hofstetter, 1998).

The sound waves produced are those already stored in a wavetable in the receiving instrument or sound card. Codes within the MIDI file tell the sound card within the computer when to turn notes on and off, what volume to use, and what instrument should make the sound. MIDI files require little bandwidth to transmit, making them a popular format for transmitting audio over the Internet (Hofstetter, 1998).

MP3

MPEG-1 audio layer-3 (MP3) compresses high-quality sound sequences into digital audio format. MP3 provides small file sizes (about one twelfth the size of the original file size) while retaining original sound quality. This results in compact audio files that download quickly without degradation of sound.

MP3 files require the use of a player, available for download from the Internet. There are multiple Web sites that provide access to MP3 music files, making them available for download. Two good sources for additional information about MP3 technology are http://MPEG.org and http://MP3.com (Hofstetter, 2001).

MULTIMEDIA FILES

Multimedia files allow the user to view digitized video and simultaneously hear accompanying sound. Some popular multimedia formats include AVI (audio/video interleave, created by Microsoft); QuickTime (created by Apple Computer); MPEG (Moving Picture Expert Group, the name of the ISO standards committee that created this format); RM or RAM (RealVideo, created by RealNetworks, Inc.); and macromedia shockwave files (Hofstetter, 1998).

AVI

The audio/video interleaved (AVI) format has been designed to conform to the Microsoft Windows resource interchange file format (RIFF) specification. AVI format files contain both sound and video. Within each file, audio frames are interleaved with video. Audio takes priority during playback, thus playing the sound track without interruption. The computer displays as many video frames as it can process during the audio playback, skipping those for which it has run out of time. This can result in the appearance of choppy motion. AVI files have a file name extension of .avi and require a special player that can be downloaded from the Internet if not included with the user's browser software (Hofstetter, 1998).

MOV/QT

QuickTime stores sound, text, animation, and video in a single file with file name extensions of .qt, .mov, or .moov. Both PCs and Macintosh computers support the QuickTime format. QuickTime files are viewed using a QuickTime player. The player may be included with browser software or downloaded from the Internet and allows the user to view and control brief multimedia sequences. (http://www.whatis.com).

MPG/MPEG/MPE

Multimedia files in MPEG format have file name extensions of .mpg, .mpe, or .mpeg. The MPEG format creates high-quality digitized video that requires the use of a MPEG player or other software program to view. There are a number of free players available for download from the Web (Hofstetter, 1998).

MPEG compresses video by eliminating redundant data in blocks of screen pixels and stores only the changes from one frame to another, instead of each entire frame. The video information is then encoded using a technique called delta-frame encoding, or DCT (Hofstetter, 1998).

There are three versions of MPEG: MPEG-1, which is a noninterlaced version used for CD-ROM playback; MPEG-2, an interlaced version designed for all-digital TV transmission; and MPEG-4, a low-bandwidth MPEG version for transmission over mobile and wireless networks as well as the Internet (Hofstetter, 2001). A fourth version, MPEG-3, was originally developed to accommodate high-definition television transmissions but was made unnecessary by the development of MPEG-2.

RM

RealVideo, created by Progressive Networks, conforms to the proposed industry standard real-time streaming protocol (RTSP) and delivers full-motion streaming video over the Internet. Streaming video may be defined as an uninterrupted sequence of moving images that have been compressed and are then transmitted over the World Wide Web. RealVideo files have file name extensions of .rm or .ram. As with streaming audio, streaming video transmissions are downloaded and displayed by the viewer as they arrive. Download and playback occur simultaneously. The user needs a player to view the video and hear the audio sound track. Players can be downloaded from the Internet if not included as part of the user's browser software (http://www.whatis.com; Hofstetter, 1998). A free RealONE media player, available from http://www.real.com, is one example of a downloadable media player.

Shockwave

Shockwave files are animations created by Macromedia's Director, Authorware, or Flash programs and included as Web page objects. These files usually have file name extensions of .dcr, .dir, or .dxr and are invoked by HTML code referencing the appropriate object file name. Shockwave supports audio and video. It runs on Windows as well as Macintosh computers. As with MPEG and QuickTime animations, Shockwave requires the use of a plug-in and player application to display and hear the files it creates (http://www.whatis.com; Hofstetter, 1998). A free Shockwave player is available for download from http://sdc.shockwave.com/shockwave/download/.

SERVER-SIDE FILES

The file types discussed earlier within this chapter reside on a Web server and are downloaded to the user's computer when accessed through the Internet. In contrast, server-side files not only reside on the Web server, they also perform some action while working from this location. Unlike other Internet file types, server-side files do not have to be downloaded to the computer user's hard drive to be used. Instead, they interact with user input, create a Web page on the fly, and then return a customized display.

Some types of server-side files interact with Web pages or a combination of Web pages, and database applications are stored on the Web server. Once connected to the Internet, the user accesses a Web page through the browser software application (Netscape, Microsoft Internet Explorer, or others). HTML coding within the Web page provides a path or map to the Web server directory where the server-side file is stored. This file is accessed and an action is performed. The action depends on the type of server-side file and the instructions it contains. It may generate a new HTML page, send information to or from an HTML form page, display an alert, take the user to a new URL (uniform resource locator, or Internet address), open a new browser window, or dictate the way Web page text is formatted and displayed.

Examples of some common server-side files include ASP or CGI, PHP, JS, and CSS.

Active Server Pages

An ASP makes use of scripting technology to create dynamic, interactive Web pages. The ASP contains HTML to define the Web page layout and embedded programming code written in a scripting language, such as visual basic script, Perlscript, or JS (Gladwin, 2001; Langley, 2001). These scripts are processed on a Microsoft Web server. As a result of this action, a Web page is created on the fly based on the user's request and data accessed from a database. This page is displayed in the browser (http://www.whatis.com).

ASP allows developers to create Web pages that may be used to interact with databases or other applications. Online forms, Web pages that contain such dynamic information as news reports and stock market data, often make use of ASP. Once the browser requests a file from the Web server, an embedded script runs, pulling up a file with an .asp extension from the Web server. The requested information is returned to the browser in the form of a customized Web page (Langley, 2001; Gladwin, 2001).

Common Gateway Interface

CGI is a standardized procedure for transmitting information between a Web server and a script. The CGI script is usually written in a programming language, such as Perl. When a user types information into a Web page form, this input is transmitted to the Web server, where it is routed through the CGI to an application program that, in turn, processes the data. The program responds to the user's input by displaying an existing Web page in the browser, generating a custom Web page on the fly, or retrieving and displaying an image or multimedia file (Hayes, 1999; Castro, 2000).

Some examples of CGI scripts used in conjunction with Web pages to retrieve database information, then return this data within the format of a Web page created on the fly, include search engines or directories, online registration forms, online order forms, and electronic full-text databases (http://www.whatis.com).

PHP

Like ASP, PHP is a script language and interpreter used primarily on Linux Web servers. The PHP script is embedded within a Web page among the HTML codes. Before the page is sent to a user who has requested it, the Web server invokes PHP to interpret and then perform the actions called for in the PHP script. An HTML file that includes a PHP script usually possesses a file name extension of .php, .php3, or .phtml (http://www.whatis.com).

JavaScript

JS is a program that can be included in HTML coding on a Web page, or as a separate script file stored on a Web server and invoked by a line of code embedded as a part of the HTML coding. JS allows the creation of interactive Web pages, validation of online forms, dynamic HTML, and the creation of customized HTML pages on the fly (Negrino & Smith, 2001).

A server-side JS file is coded as a separate text file, saved with a file extension of .js, and stored on the Web server. A corresponding line of HTML code that tells the browser the URL of the .js file is embedded in the Web page's HTML coding. When that Web page is accessed, the JS file is retrieved from the server and runs inside the Web browser (Negrino & Smith, 2001).

Cascading Style Sheets

Like JS, CSS coding can either be included in a Web page's HTML coding or saved as a separate text file that is stored on the Web server and invoked by a line of HTML code within a Web page. A server-side CSS is saved with a file name extension of .css. A style sheet allows Web page designers to specify the look and format of Web pages; set font face, style, size, and color; and define precise positioning of such elements as images or text, or both (Burns, 1999).

FINDING AND DOWNLOADING FILES FROM THE INTERNET

Some search utilities, such as Google (http://www.google.com/), Alta Vista (http://www.altavista.com/), and HotBot (http://hotbot.lycos.com/), provide the ability to limit searches according to format (images, video, MP3, audio, etc.). Users may use these utilities and other tools to find files in a specific format. FTP archives are another source of various Internet files. Some general-purpose FTP archives include the following resources: the University of Illinois at Urbana-Champaign UIArchive, http://uiarchive.uiuc.edu/; the Washington University in St. Louis Wuarchive, http://wuarchive.wustl.edu/; the University of Vassa, Finland, Garbo Anonymous FTP Archive, http://garbo.uwasa.fi/; and the Monash University Monash Nihongo FTP Archive, http://ftp.monash.edu.au/pub/nihongo/00INDEX.html.

Saving Internet Files

Text within a Web page may be saved by using the browser's Save As function. Once the Web page is displayed in the browser, the user chooses File and then Save As and then selects Save As Type: Text (Figures 1 and 2).

To save an image found on the Internet, place the mouse cursor directly on the image and right click (PC) or click (Macintosh). Choose Save Image As, rename or accept the existing file name, and indicate the directory location where you wish to save the file (Figures 3 and 4) (Ackermann & Hartman, 2001).

Files may also be exchanged over the Internet using file transfer protocol (FTP) software. Unlike HTTP (hypertext transfer protocol), FTP allows files to be transferred over the Internet from a Web server to the hard drive of the user's computer without first retrieving and then opening the file. Using FTP software configured to access the host computer's address, a user can locate and transfer one or multiple files. When the user knows the exact name and location of a file, FTP can be a quick and efficient way to copy files from one computer to another (Ackermann & Hartman, 2001).

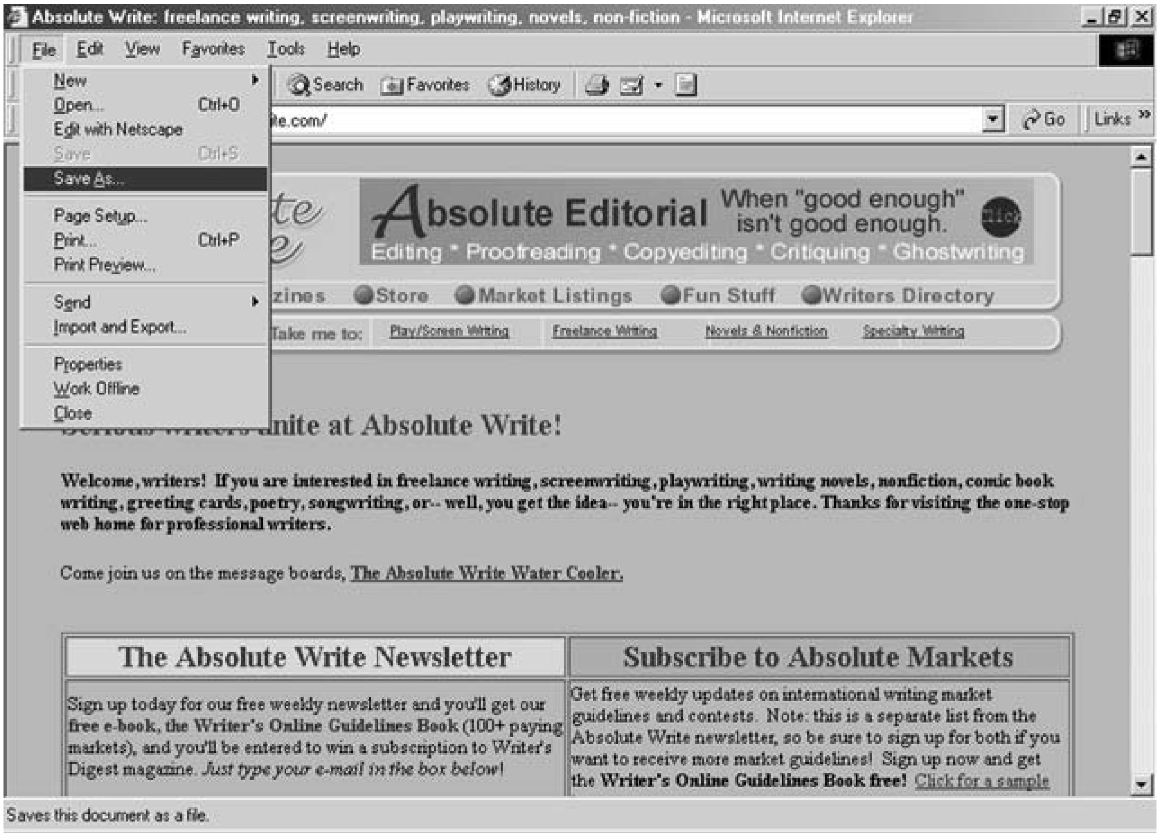
What is FTP?

The name FTP is an acronym for file transfer protocol. Downloading is what it is called when a user copies a file from someone else's (however remote) computer to his/her own. When the user sends a file from his/her computer to someone else's, this is defined as uploading the file.

In FTP terminology, the user's computer is called the local host and the other computer is called the remote host. It does not matter where the two computers are located, how they are connected, or even whether they use the same operating system. Using FTP, the two computers can, through the Internet, exchange files.

Anonymous FTP

Collections of files are made available for public downloading by storing them in a public FTP directory located


Absolute Write: freelance writing, screenwriting, playwriting, novels, non-fiction - Microsoft Internet Explorer
File Edit View Favorites Tools Help
New
Open... Ctrl+O
Edit with Netscape
Save Ctrl+S
Save As...
Page Setup...
Print... Ctrl+P
Print Preview...
Send
Import and Export...
Properties
Work Offline
Close
Search Favorites History
te.com/
Go Links
Absolute Editorial
When "good enough" isn't good enough.
Editing * Proofreading * Copyediting * Critiquing * Ghostwriting
zines Store Market Listings Fun Stuff Writers Directory
Take me to: Play/Screen Writing Freelance Writing Novels & Nonfiction Specialty Writing

nite at Absolute Write!

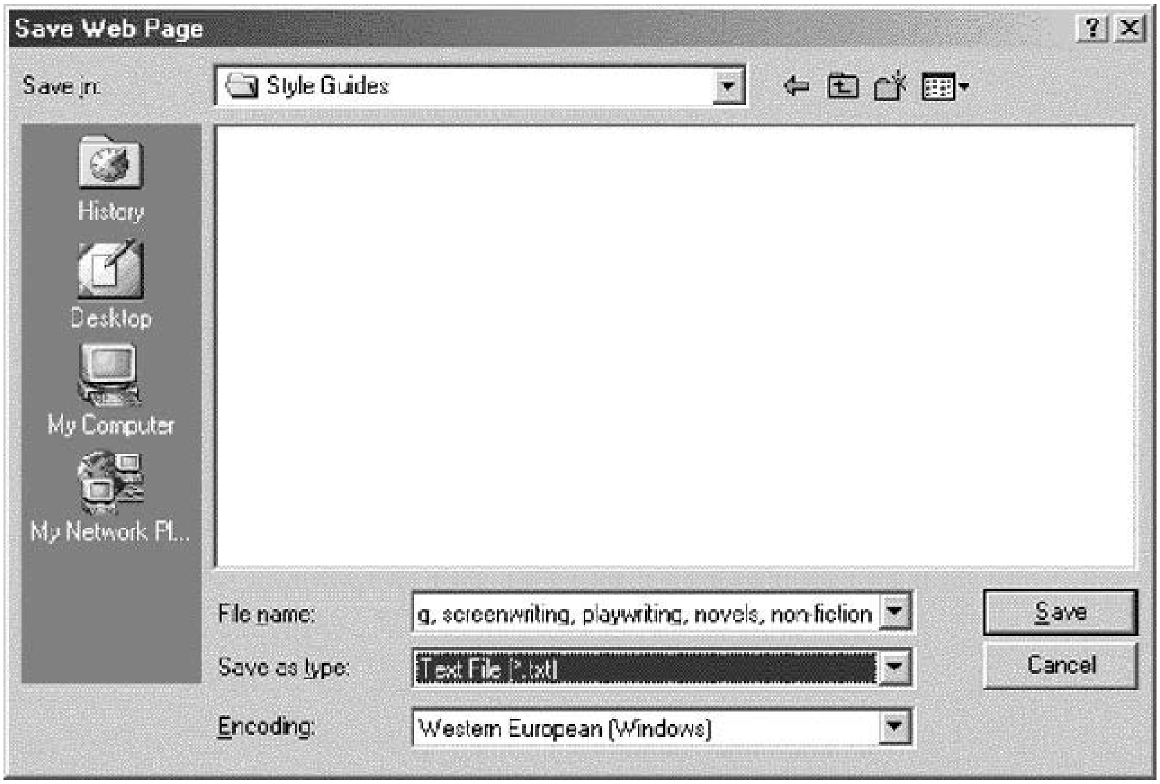
Welcome, writers! If you are interested in freelance writing, screenwriting, playwriting, writing novels, nonfiction, comic book writing, greeting cards, poetry, songwriting, or-- well, you get the idea-- you're in the right place. Thanks for visiting the one-stop web home for professional writers.

Come join us on the message boards, The Absolute Write Water Cooler.

The Absolute Write Newsletter	Subscribe to Absolute Markets
Sign up today for our free weekly newsletter and you'll get our free e-book, the **Writer's Online Guidelines Book** (100+ paying markets), and you'll be entered to win a subscription to Writer's Digest magazine. *Just type your e-mail in the box below!*	Get free weekly updates on international writing market guidelines and contests. Note: this is a separate list from the Absolute Write newsletter, so be sure to sign up for both if you want to receive more market guidelines! Sign up now and get the **Writer's Online Guidelines Book free!** Click for a sample

Saves this document as a file.

Figure 1: Illustration of browser File, Save As drop down menu.



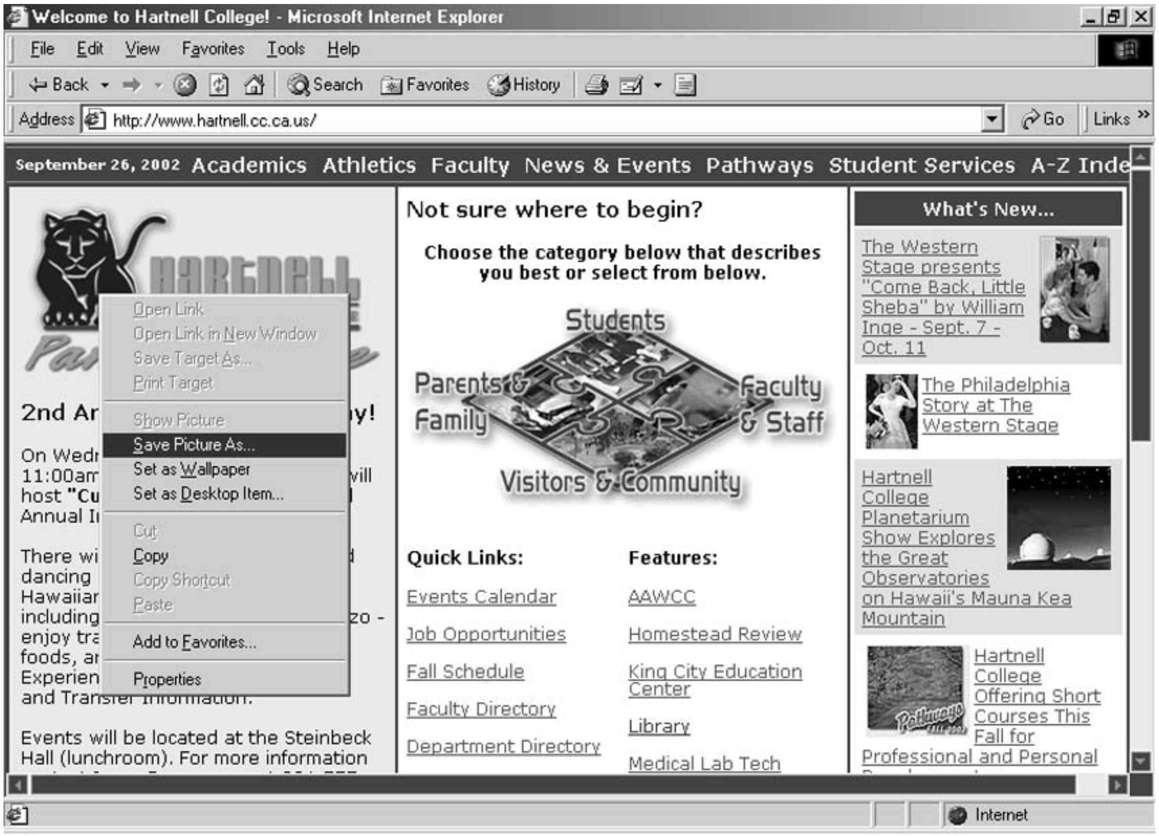
Save in: Style Guides

File name: g, screenwriting, playwriting, novels, non-fiction

Save as type: Text File (*.txt)

Encoding: Western European (Windows)

Figure 2: Illustration of file type options within display window during save operation.



Address: http://www.hartnell.cc.ca.us/

2nd Ar… …y!

On Wedr… 11:00am… …vill host "Cu… Annual I…

There wi… dancing… Hawaiiar… including… …zo - enjoy tra… foods, ar… Experien… and Transfer Information.

Events will be located at the Steinbeck Hall (lunchroom). For more information

Not sure where to begin?

Choose the category below that describes you best or select from below.

Quick Links:

- Events Calendar
- Job Opportunities
- Fall Schedule
- Faculty Directory
- Department Directory

Features:

- AAWCC
- Homestead Review
- King City Education Center
- Library
- Medical Lab Tech

What's New...

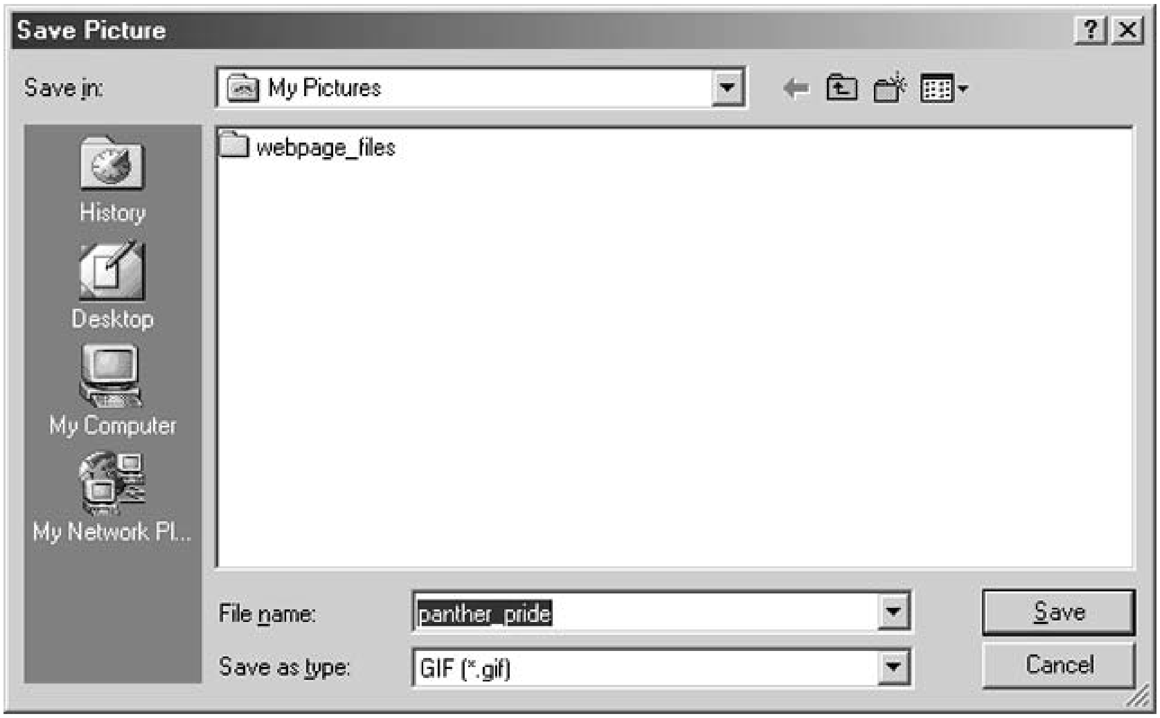
The Western Stage presents "Come Back, Little Sheba" by William Inge - Sept. 7 - Oct. 11

The Philadelphia Story at The Western Stage

Hartnell College Planetarium Show Explores the Great Observatories on Hawaii's Mauna Kea Mountain

Hartnell College Offering Short Courses This Fall for Professional and Personal

Figure 3: Illustration of drop down menu screen display during save operation.



Save in: My Pictures

webpage_files

File name: panther_pride

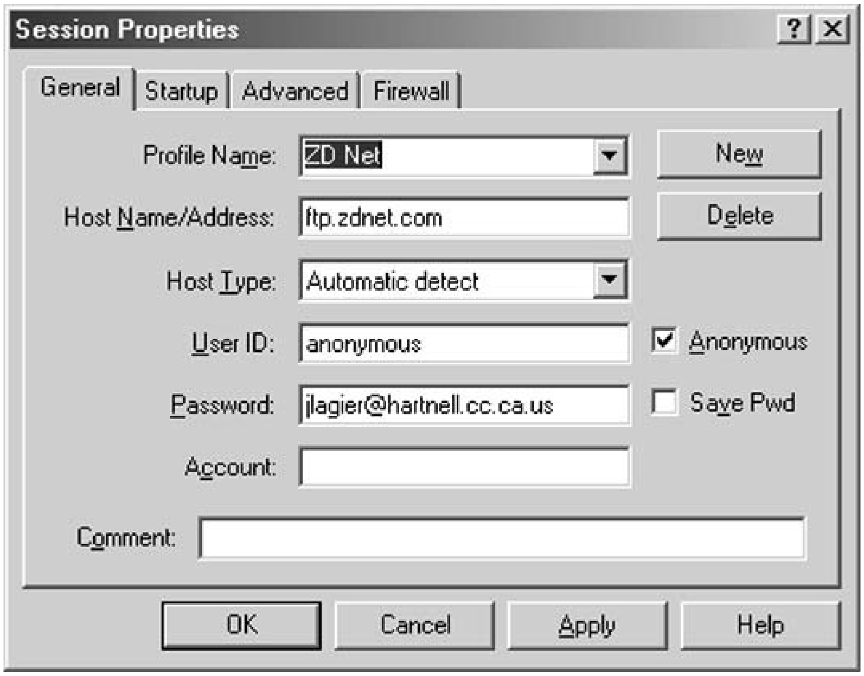
Save as type: GIF (*.gif)

Figure 4: Illustration of file type options within display window during save operation.

on an FTP server. Anonymous FTP is a facility that lets users log onto remote hosts and download files by using the user identification (log-on) of anonymous. By using a log-on of anonymous, the need to supply a regular password is eliminated. Instead, the user supplies his or her e-mail address (Figure 5) (Lehnert, 1998).

Transferring Files

All FTP clients work more or less the same way. The screen display window is be divided into two sections, one on the left and one on the right. The left section displays information about the directory structure and files on the local computer. The right section shows the



Profile Name: ZD Net

Host Name/Address: ftp.zdnet.com

Host Type: Automatic detect

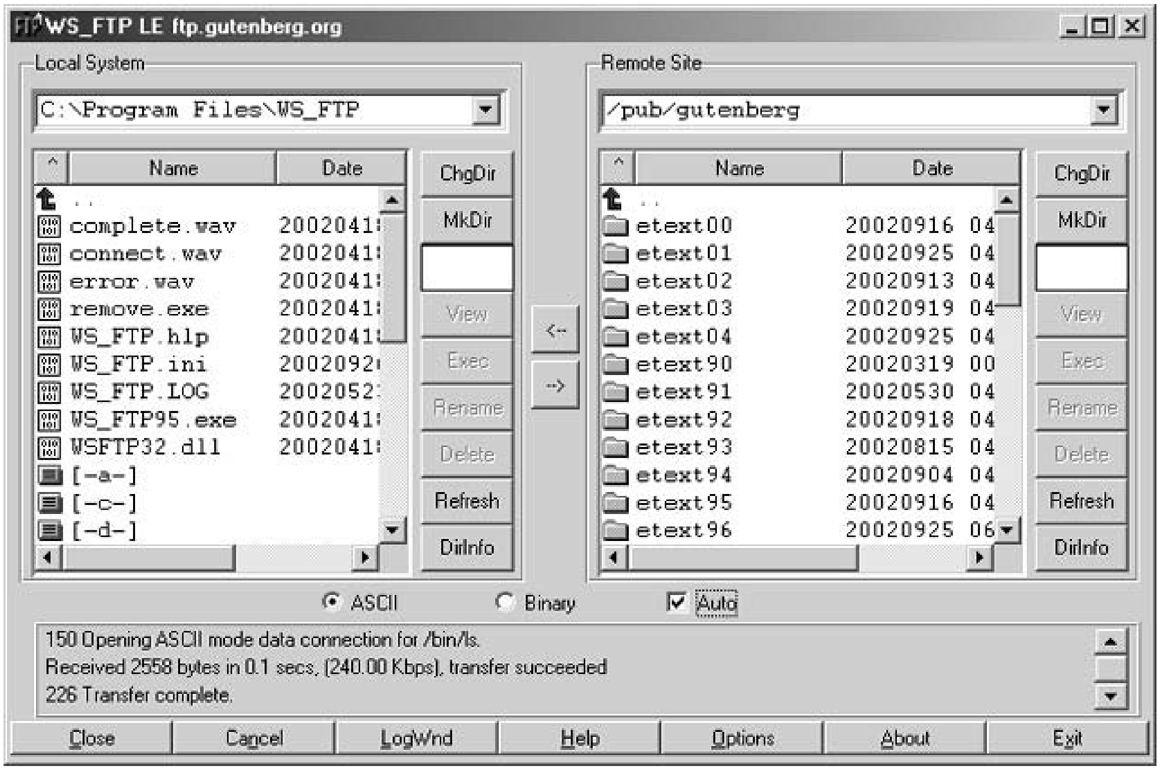
User ID: anonymous ☑ Anonymous

Password: jlagier@hartnell.cc.ca.us ☐ Save Pwd

Account:

Comment:

Figure 5: Illustration of ftp client session properties screen display.



C:\Program Files\WS_FTP

```
..
complete.wav    2002041
connect.wav     2002041
error.wav       2002041
remove.exe      2002041
WS_FTP.hlp      2002041
WS_FTP.ini      2002092
WS_FTP.LOG      2002052
WS_FTP95.exe    2002041
WSFTP32.dll     2002041
[-a-]
[-c-]
[-d-]
```

/pub/gutenberg

```
..
etext00    20020916 04
etext01    20020925 04
etext02    20020913 04
etext03    20020919 04
etext04    20020925 04
etext90    20020319 00
etext91    20020530 04
etext92    20020918 04
etext93    20020815 04
etext94    20020904 04
etext95    20020916 04
etext96    20020925 06
```

● ASCII ○ Binary ☑ Auto

```
150 Opening ASCII mode data connection for /bin/ls.
Received 2558 bytes in 0.1 secs, (240.00 Kbps), transfer succeeded
226 Transfer complete.
```

Figure 6: Illustration of ftp client program screen display.

names of the directories and files on the remote computer (Figure 6).

The general process of accessing anonymous FTP resources involves the following steps:

1. Establish a connection to the remote host.
2. Navigate to the directory on the remote host that contains the desired file.
3. Set the desired transfer options. Usually an ASCII transfer is used for text files. A binary transfer is used to transfer software or graphic image files. This transfer mode simply moves the file without attempting to open or change it.
4. Select the file to be transferred.
5. Designate the desired directory on the local computer in which to store the transferred file.
6. Initiate the file transfer (Hofstetter, 1998; Ackermann & Hartman, 2001).

Downloading Files and Software

At times, users may wish to download extremely large files or entire software programs. To reduce the cost and space of storage and download time, large files are frequently stored in a compressed format. Such files should be treated as binary files (i.e., files that require translation by a software program or hardware processor). The problem is that getting the file transferred from the remote host to the user's local computer is only half the process. Once it is there, the user will need to extract or unzip the file to make it usable (Ackermann & Hartman, 2001).

When using FTP, the user may want to move an entire directory or collection of directories. At this time, there is no standard command or effective means of accomplishing this in a single step. Instead, the most common practice is to move multiple files from an FTP archive or back up a selection of files into an aggregate file, which then can be moved (Ackermann & Hartman, 2001).

Although current technology makes it possible for users to easily download files from the Internet and save them on the hard drives of their own computers, it is important to observe copyright restrictions. Only the individual or entity that retains ownership can grant the right to copy or distribute materials, including those found on the World Wide Web. Detailed information describing copyright restrictions and fair use doctrine is provided elsewhere in this encyclopedia.

ONLINE RESOURCES

Browsers

Netscape Browser Central, http://browsers.netscape.com/browsers/main.tmpl

Microsoft Internet Explorer, http://www.microsoft.com/windows/ie/default.asp

Images

Google Image Search, http://images.google.com

Yahoo! Computers and Internet: Graphics, http://dir.yahoo.com/Computers_and_Internet/Graphics/

MP3

MPEG.org, http://www.mpeg.org/MPEG/index.html

MP3.com, http://www.mp3.com/

Multimedia Players

Apple QuickTime Player, http://www.apple.com/quicktime/products/qt/

RealOne Player, http://www.real.com/realoneplayer.html?src = 011204help,020708r1choice_c2

Shockwave Player, http://www.macromedia.com/software/shockwaveplayer/index.html

Software (Helpers and Plug-ins)

CWS Apps, http://cws.internet.com/browsers.html

Yahoo Plug-ins, http://dir.yahoo.com/Computers_and_Internet/software/internet/world_wide_web/browsers/plug_ins/

Web Developers Virtual Library Plug-ins, http://www.wdvl.com/Software/Plugins/

More Information About File Formats

Whatis.com, http://whatis.techtarget.com/definition/0,,sid9_gci834191,00.html

File Extensions, Formats, and Utilities, http://www.stack.com/

GLOSSARY

Active server pages A Web or HTML page that includes an embedded program called a script. A Microsoft Web server processes the script, creating a Web page on the fly.

Bitmap A graphic that is pixel based, defining display space and color for each pixel within an image. Both GIF and JPG files contain bitmaps.

Browser or Web browser Software that permits the user to traverse the Internet, navigating between Web sites, accessing and downloading files. Examples include Microsoft Internet Explorer and Netscape Navigator.

Cascading style sheets Specify the look and format of Web pages; set font face, style, size, and color; and define precise positioning of images or text, or both. Style sheet coding can be included within a Web page's HTML coding, or it can be a stored on the Web server as a separate file.

Common gateway interface (CGI) A server-side script that transmits information between a Web page and an application program. Search engines and online forms use CGI scripts to communicate with the server, returning information to the user by means of Web pages created on the fly.

File name extension A two- to four-character string that is appended to a file name following a period. The file name extension can be used to associate a software program with the file. For example, a file name of *article.txt* indicates this file is in text format.

Format A preestablished layout for information. All data is stored in some format. A program that recognizes that format will be able to retrieve, read, and process the stored information.

FTP (File Transfer Protocol) The method used to transfer files over the Internet.

GIF (Graphics Interchange Format) The most commonly used image format on the Web.

Helper application A program that works in conjunction with the Web browser to activate and play a multimedia file. Two examples of helper applications are RealPlayer or QuickTime.

HTML (Hypertext Markup Language) The markup code used within a text format file to create Web page files.

Hypertext Text on a Web page that has been linked to another Web page, a location in the current Web page, or an image, audio, video, or multimedia file. Clicking on the hypertext activates the resource to which it is connected. Hypertext files have file name extensions of .htm or .html.

JavaScript A scripting language developed by Netscape Communications that adds interactivity to Web pages and supports dynamic HTML or DHTML. Examples include alerts, pop-up windows, image swapping, and scrolling text banners.

JPG or JPEG (Joint Photographic Experts Group) File format used for full-color images on the Internet.

MID or MIDI (Musical Instrument Digital Interface) File format used to transmit music files over the Internet.

MPG or MPEG (Motion Pictures Expert Group) File format for digital video files named after the ISO (International Organization for Standardization) standards committee. There are four versions of MPEG. MPEG-1 is designed for playback from CD-ROMS. MPEG-2 is used for all-digital, broadcast-quality television transmission. MPEG-3 was to be used for high-definition TV but the development of MPEG-2 made it unnecessary. MPEG-4 permits the transmission of movies using wireless or mobile communications.

MP3 A method and format for compressing audio files.

PHP A script language and interpreter used primarily on Linux Web servers.

Plug-in A software application used in conjunction with a Web browser to view or display certain file formats. Examples of plug-ins include Shockwave and Adobe Acrobat Reader.

Streaming media Sound or video transmitted over the Internet in such a way that a portion of the file downloads into a buffer and then starts playing as the remainder of the file is downloaded. Download and playback occur simultaneously in an uninterrupted stream.

Text file A file containing readable characters with no formatting commands such as underlining, bold face, or italics. The file name extension for a text file is .txt.

URL (Uniform Resource Locator, or Internet address) Examples are http://www.google.com and ftp://ibiblio.org/pub/docs/books/gutenberg/.

CROSS REFERENCES

See *Cascading Style Sheets (CSS); Common Gateway Interface (CGI) Scripts; Downloading from the Internet; HTML/XHTML (HyperText Markup Language/Extensible HyperText Markup Language); JavaScript; Multimedia.*

REFERENCES

Ackerman, E., & Hartman, K. (2001). *Internet and Web essentials: What you need to know.* Wilsonville, OR: Franklin, Beedle and Associates.

Adobe Acrobat Reader (2002). Retrieved July 12, 2002, from http://www.adobe.com/products/acrobat/readstep2.html.

Burns, J. (1999). *HTML goodies*. Indianapolis, IN: Que.

Castro, E. (2000). *HTML 4 for the World Wide Web*. Berkeley, CA: Peachpit Press.

Gladwin, L.C. (2001). Active server pages. *Computerworld, 35* (12), 64.

Hayes, F. (1999). Common gateway interface. *Computerworld*, *33* (29), 74.

Hofstetter, F.T. (1998). *Internet Literacy*. Boston: Irwin McGraw-Hill.

Hofstetter, F.T. (2001). *Multimedia literacy* (3rd ed). Boston: Irwin McGraw-Hill.

Langley, N. (2001). Web page construction made easy. *Computer Weekly, May 31,* 70.

Lehnert, W.G. (1998). *Internet 101: A beginner's guide to the Internet and the World Wide Web*. Reading, MA: Addison-Wesley.

Miller, J. (1997). What's in a name? *Technology Connection, 4* (7), 20–22.

Negrino, T., & Smith, D. (2001). *Visual Quickstart guide: JavaScript for the World Wide Web* (4th ed). Berkeley, CA: Peachpit Press.

Web Developers Virtual Library (2002). *Graphics formats for the World Wide Web*. Retrieved July 12, 2002, from http://www.wdvl.com/Graphics/Formats/.

Weinman, L. (1999). *Designing Web graphics.3: How to prepare images and media for the Web*. Indianapolis, IN: New Riders Publishing.

Whatis.com. (2002). Retrieved July 12, 2002, from http://www.whatis.com.

Firewalls

James E. Goldman, *Purdue University*

INTRODUCTION

When an organization or individual links to the Internet, a two-way access point out of and in to their information systems is created. To prevent unauthorized activities between the Internet and the private network, a specialized hardware, software, or software–hardware combination known as a firewall is often deployed.

Firewall software often runs on a dedicated server between the Internet and the protected network. Firmware-based firewalls and single-purpose dedicated firewall appliances are situated in a similar location on a network and provide similar functionality to the software-based firewall. All network traffic entering the firewall is examined, and possibly filtered, to ensure that only authorized activities take place. This process may be limited to verifying authorized access requested files or services, or it may delve more deeply into content, location, time, date, day of week, participants, or other criteria of interest. Firewalls usually provide a layer of isolation between the inside, sometimes referred to as the "clean" network, and the outside or "dirty" network. They are also used, although less frequently, to separate multiple subnetworks so as to control interactions between them.

A common underlying assumption in such a design scenario is that all of the threats come from the outside network or the Internet, but many modern firewalls provide protection against insiders acting inappropriately and against accidental harm that could result from internal configuration errors, viruses, or experimental implementations. Research consistently indicates that 70–80% of malicious activity originates from insiders who would normally have access to systems both inside and outside of the firewall. In addition, outside threats may be able to circumvent the firewall entirely if dial-up modem access remains uncontrolled or unmonitored, if radio-LANs (local area networks) or similar wireless technology is used, or if other methods can be used to co-opt an insider, to subvert the integrity of systems within the firewall, or to otherwise bypass the firewall as a route for communications. Incorrectly implemented firewalls can exacerbate this situation by creating new, and sometimes undetected, security holes or by creating such an impediment to legitimate uses that insiders subvert its mechanisms intentionally. It is often said that an incorrectly configured and maintained firewall is worse than no firewall at all because it gives a false sense of security.

Advantages

When properly configured and monitored, firewalls can be an effective way to protect network and information resources. Firewalls are often used to reduce the costs associated with protecting larger numbers of computers that are located inside it. Firewalls provide a control point that can be used for other protective purposes.

Disadvantages

- Firewalls can be complex devices with complicated rule sets.
- Firewalls must be configured, managed, and monitored by properly trained personnel
- A misconfigured firewall gives the illusion of security.
- Even a properly configured firewall is not sufficient or complete as a perimeter security solution.
- The processing of packets by a firewall inevitably introduces latency which, in some cases, can be significant.
- Firewalls often interfere with network applications that may expect direct connections to end user workstations such as Voice over Internet Protocol (IP) phones and VPNs (virtual private networks) as well as some collaborative tools such as instant messenger, and video and audio conferencing software.

FIREWALL BASICS

A given firewall may, but does not necessarily, offer any or all of the following functions.

Address Filtering

This function can, and perhaps should, be performed by a device other than the firewall, such as a border router. Based on address filter tables containing allowed or disallowed individual or groups (subnets) or addresses, address filtering blocks all traffic that contains either disallowed source or destination addresses. Address filtering by itself may not provide adequate protection because it is too large a granularity for some protective requirements.

Port Filtering

Port filtering goes beyond address filtering, examining to which function or program a given packet applies. For example, file transfer protocol (FTP) typically uses ports 20 and 21, inbound electronic mail (simple mail transfer protocol; SMTP) typically uses port 25, and inbound nonencrypted Web traffic (hypertext transfer protocol; HTTP) typically uses port 80. Attacks are often targeted and designed for specific ports. Port filtering can be used to ensure that all ports are disabled except those that must remain open and active to support programs and protocols required by the owners of the inside network. Static port filtering leaves authorized ports open to all traffic all the time, whereas dynamic port filtering opens and closes portions of protocols associated with authorized ports over time as required for the specifics of the protocols.

Domain Filtering

Domain filtering applies to outbound traffic headed through the firewall to the Internet. Domain filtering can block traffic with domains that are not authorized for communications or can be designed to permit exchanges only with authorized "outside" domains.

Network Address Translation

When organizations connect to the Internet, the addresses they send out must be globally unique. In most cases, an organization's Internet service provider specifies these globally unique addresses. Most organizations have relatively few globally unique public addresses compared with the actual number of network nodes on their entire network.

The Internet Assigned Numbers Authority has set aside the following private address ranges for use by private networks:

10.0.0.0 to 10.255.255.255
172.16.0.0 to 172. 31.255.255
192.168.0.0 to 192.168.255.255

Traffic using any of these "private" addresses must remain on the organization's private network to interoperate properly with the rest of the Internet. Because anyone is welcome to use these address ranges, they are not globally unique and therefore cannot be used reliably over the Internet. Computers on a network using the private IP address space can still send and receive traffic to and from the Internet by using network address translation (NAT). An added benefit of NAT is that the organization's private network is not as readily visible from the Internet.

All of the workstations on a private network can share a single or small number of globally unique assigned public IP address(es) because NAT mechanism maintains a table that provides for translation between internal addresses and ports and external addresses and ports. These addresses and port numbers are generally configured so as to not to conflict with commonly assigned transmission control protocol (TCP) port numbers. The combination of the shared public IP address and a port number that is translated into internal address and port numbers via the translation table allows computers on the private network to communicate with the Internet through the firewall. NAT can also run on routers, dedicated servers, or other similar devices, and "gateway" computers often provide a similar function.

Data Inspection

The primary role of many firewalls is to inspect the data passing through it using a set of rules that define what is and is not allowed through the firewall and then to act appropriately on packets that meet required criteria. The various types of firewalls, described in the next section, differ primarily in what portions of the overall data packet are inspected, the types of inspections that can be done, and what sorts of actions are taken with respect to that data. Among the data elements that are commonly inspected are the following:

- IP address
- TCP port number
- User Datagram Protocol port number
- Data field contents
- Contents of specific protocol payloads, such as HTTP, to filter out certain classes of traffic (e.g., streaming video, voice, music, etc.), access requests to restricted Web sites, information with specific markings (e.g., proprietary information), content with certain words or page names

Virus Scanning and Intrusion Detection

Some firewalls also offer functionality such as virus scanning and intrusion detection. Advantages of such firewalls include the following:

- "One-stop shopping" for a wide range of requirements,
- Reduced overhead from centralization of services, and
- Reduced training and maintenance.

Disadvantages include the following:

- Increased processing load requirements,
- Single point of failure for security devices,
- Increased device complexity, and
- Potential for reduced performance over custom subsolutions.

FIREWALL TYPES

Another difficulty with firewalls is that there are no standards for firewall functionality, architecture, or interoperability. As a result, users must often be aware of how firewalls work to use them, and owners must be aware of these issues to evaluate potential firewall technology purchases. Firewall functionality and architectures are explained in the next few sections.

Bastion Host

Many firewalls, whether software or hardware based, include a bastion host—a specially hardened server or a trusted system designed so that the functionality of the device cannot be compromised by attacking vulnerabilities in the underlying operating system or software over which its software runs. Specifically, the bastion host employs a secure version of the operating system with the most recent patches, security updates, and minimum number of applications to avoid known and unknown vulnerabilities.

Packet Filtering Firewalls

Every packet of data on the Internet can be identified by a source address normally associated with the computer that issued the message and the destination address normally associated with the computer to which the message is bound. These addresses are included in a portion of the packet called the header.

A packet filter can be used to examine the source and destination address of every packet. Network access devices known as routers are among the commonly used devices capable of filtering data packets. Filter tables are lists of addresses with data packets and embedded messages that are either allowed or prohibited from proceeding through the firewall. Filter tables may also limit the access of certain IP addresses to certain services and subservices. This is how anonymous FTP users are restricted to only certain information resources. It takes time for a firewall server to examine the addresses of each packet and compare those addresses to filter table entries. This filtering time introduces latency to the overall transmission time and may create a bottleneck to high volumes of traffic. A filtering program that only examines source and destination addresses and determines access based on the entries in a filter table is known as a port-level filter or network-level filter or packet filter. Hardware implementations of such filters are often used to provide low latency and high throughput.

Packet filter gateways can be implemented on routers. This means that an existing piece of technology can be used for dual purposes. Maintaining filter tables and access rules on multiple routers is not a simple task, and packet filtering of this sort is limited in what it can accomplish because it only examines certain areas of each packet. Dedicated packet-filtering firewalls are usually easier to configure and require less in-depth knowledge of protocols to be filtered or examined. One easy way that many packet filters can be defeated by attackers is a technique known as IP spoofing. Because these simple packet filters make all filtering decisions based on IP source and destination addresses, an attacker can often create a packet designed to appear to come from an authorized or trusted IP address, which will then pass through such a firewall unimpeded.

Application Gateways

Application-level filters, sometimes called assured pipelines, application gateways, or proxy servers, go beyond port-level filters in their attempts to control packet flows. Whereas port-level filters determine the legitimacy of the IP addresses and ports within packets, application-level filters are intended to provide increased assurance of the validity of packet content in context. Application-level filters typically examine the entire request for data rather than just the source and destination addresses. Controlled files can be marked as such, and application-level filters can be designed to prevent those files from being transferred, even within packets authorized by port-level filters. Of course, this increased level of scrutiny comes at the cost of a slower or more expensive firewall.

Certain application-level protocols commands that are typically used for probing or attacking systems can be identified, trapped, and removed. For example, SMTP is an e-mail interoperability protocol that is a member of the TCP/IP family and used widely over the Internet. It is often used to mask attacks or intrusions. Multipurpose internet mail extension (MIME) is another method that is often used to hide or encapsulate malicious code such as Java applets or ActiveX components. Other application protocols that may require monitoring include but are not limited to World Wide Web protocols such as HTTP, telnet, ftp, gopher, and Real Audio. Each of these application protocols may require its own proxy, and each application-specific proxy must be designed to be intimately familiar with the commands within each application that will need to be trapped and examined. For example an SMTP proxy should be able to filter SMTP packets according to e-mail content, message length, and type of attachments. A given application gateway may not include proxies for all potential application layer protocols.

Circuit-Level Proxies

Circuit-level proxies or circuit-level gateways provide proxy services for transport layer protocols such as TCP. Socks, an example of such a proxy server, creates proxy data channels to application servers on behalf of the application client. Socks uniquely identifies and keeps track of individual connections between the client and server ends of an application communication over a network. Like other proxy servers, both a client and server portion of the Socks proxy are required to create the Socks tunnel. Some Web browsers have the client portion of Socks included, whereas the server portion can be added as an additional application to a server functioning as a proxy server. The Socks server would be located inside an organization's firewall and can block or allow connection requests, based on the requested Internet destination, TCP port ID, or user identification. Once Socks approves and establishes the connection through the proxy server, it does not care which protocols flow through the established connection. This is in contrast to other more protocol-specific proxies such as Web proxy, which only allows HTTP to be transported, or WinSock Proxy, which only allows Windows application protocols to be transported.

Because all data goes through Socks, it can audit, screen, and filter all traffic in between the application client and server. Socks can control traffic by disabling or enabling communication according to TCP port numbers. Socks4 allowed outgoing firewall applications, whereas Socks5 supports both incoming and outgoing firewall applications, as well as authentication.

The key negative characteristic of Socks is that applications must be "socksified" to communicate with the Socks protocol and server. In the case of Socks4, this meant that local applications had to be recompiled using a Socks library that replaced its normal library functions. However, with Socks5, a launcher is employed which avoids "socksification" and recompilation of client programs that in most cases do not natively support Socks. Socks5 also uses a private routing table and hides internal network addresses from outside networks.

Application gateways are concerned with what services or applications a message is requesting in addition to who is making that request. Connections between requesting clients and service providing servers are only created after the application gateway is satisfied as to the legitimacy of the request. Even once the legitimacy of the request has been established, only proxy clients and servers actually communicate with each other. A gateway firewall does not allow actual internal IP addresses or names to be transported to the external nonsecure network, except as this information is contained within content that the proxy does not control. To the external network, the proxy application on the firewall appears to be the actual source or destination, as the case may be.

Trusted Gateway

A trusted gateway or trusted application gateway seeks to relieve all the reliance on the application gateway for all communication, both inbound and outbound. In a trusted gateway, certain applications are identified as trusted and are able to bypass the application gateway entirely and are able to establish connections directly rather than be executed by proxy. In this way, outside users can access information servers and Web servers without tying up the proxy applications on the application gateway. These servers are typically placed in a demilitarized zone (DMZ) so that any failures in the application servers will grant only limited additional access to other systems.

Proxies are also capable of approving or denying connections based on directionality. Users may be allowed to upload but not download files. Some application-level gateways have the ability to encrypt communications over these established connections. The level of difficulty associated with configuring application-level gateways versus router-based packet filters is debatable. Router-based gateways tend to require a more intimate knowledge of protocol behavior, whereas application-level gateways deal predominantly at the application layer of the protocol stack. Proxies tend to introduce increased latency compared with port-level filtering. The key weaknesses of an application-level gateway is their inability to detect embedded malicious code such as Trojan horse programs or macro viruses and the requirement of more complex and resource intensive operation than lower level filters.

Stateful Firewalls

Rather than simply examining packets individually without the context of previously transmitted packets from the same source, a stateful firewall stores information about past activities and uses this information to test future packets attempting to pass through. Stateful firewalls typically review the same packet information as normal simple packet filtering firewalls such as source address, destination address, protocol, port, and flags; however, they also record this information in a connection state table before sending the packet on. This table will have an entry for each valid connection established over a particular reference frame.

Some stateful firewalls keep sequence number information to validate packets even further, so as to protect against some session hijacking attacks. As each packet arrives to a stateful firewall, it is checked against the connection table to determine whether it is part of an existing connection. The source address, destination address, source port, and destination port of the new packet must match the table entry. If the communication has already been authorized, there is no need to authorize it again and the packet is passed. If a packet can be confirmed to belong to an established connection, it is much less costly to send it on its way based on the connection table rather than reexamine the entire firewall rule set. This makes a stateful firewall faster then a simple packet filtering firewall for certain types of traffic patterns, since the packet filtering firewall treats each connection as a new connection. Additionally, because there is a 'history' in the state tables, flags can be analyzed to ensure the proper sequence as in the TCP connection handshake, and the stateful firewall can drop or return packets that are clearly not a genuine response to a request.

As a result, stateful firewalls are better able to prevent some sorts of session hijacking and man-in-the-middle attacks. Session hijacking is very similar to IP spoofing, described earlier, except that it hijacks the session by sending forged acknowledgment (ACK) packets so that the victimized computer still thinks it is talking to the legitimate intended recipient of the session. A man-in-the-middle attack could be seen as two simultaneous session hijackings. In this scenario, the hacker hijacks both sides of the session and keeps both sides active and appearing to talk directly to each other when, in fact, the man-in-the middle is examining and potentially modifying any packet that flows in either direction of the session.

Internal Firewalls

Not all threats to a network are perpetrated from the Internet by anonymous attackers, and firewalls are not a stand-alone, technology-based quick fix for network security. In response to the reality that most losses due to computer crime involve someone with inside access, internal

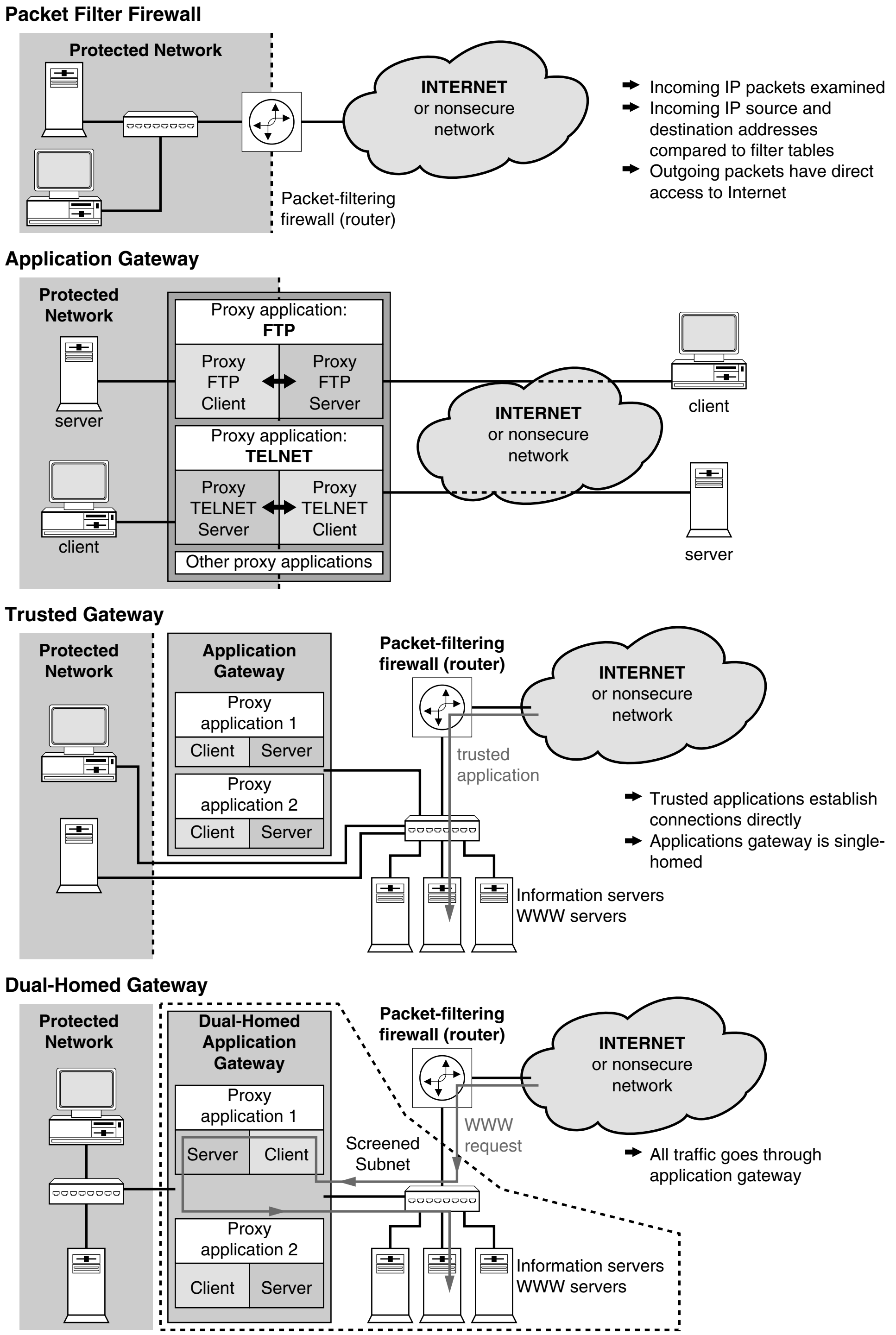


Figure 1: Firewall types.

firewalls have been applied with increasing frequency. Internal firewalls include filters that work on the data link, network, and application layers to examine communications that occur only within internal networks. Internal firewalls also act as access control mechanisms, denying access to applications for which a user does not have specific access approval. To ensure the confidentiality and integrity of private information, encryption and authentication may also be supported by firewalls, even during internal communications. Figure 1 illustrates some of the aforementioned types of firewalls.

ENTERPRISE FIREWALL ARCHITECTURES

The previous section described different approaches to firewall architecture on an individual basis, but key decisions remain to be made regarding the number and location of these firewalls in relation to the Internet and a corporation's public and private information resources. Each of the alternative enterprise firewall architectures explored in this section attempt to segregate three distinct networks or risk domains:

1. The Internet: contains legitimate customers and business partners as well as hackers
2. The DMZ, otherwise known as the screened subnet, neutral zone, or external private network: contains Web servers and mail servers
3. The internal private network, otherwise known as the secure network or intranet: contains valuable corporate information

Packet Filtering Routers

Packet filtering routers or border routers are often the first device that faces the Internet from an organization network perspective. These packet filtering routers first remove all types of traffic that should not even be passed to the firewalls. This process is typically fast and inexpensive because it involves only simple processes that are implemented in relatively low-cost hardware. Examples of removed packets include packet fragments, packets with abnormally set flags, time to live (TTL), abnormal packet length, or Internet control message protocol (ICMP) packets, all of which could potentially be used for attacks or to exploit known vulnerabilities and packet from or to unauthorized addresses and ports.

Dual-Homed Host Firewalls

In a dual-homed gateway or dual-homed host firewalls scenario, the application gateway is physically connected to the private secure network and the packet filtering router is connected to the nonsecure network or the Internet. Between the application gateway and the packet filter router is an area known as the screened subnet or DMZ. Also attached to this screened subnet are information servers, Web servers, or other servers that the company may wish to make available to outside users. All outside traffic still goes through the application gateway first, however, and then to the information servers. TCP/IP forwarding is disabled, and access to the private network is only available through one of the installed proxies. Remote logins, if they are allowed at all, may only allowed to a gateway host.

Screened Host Firewalls

An alternative to the dual-homed gateway that seeks to relieve the reliance on the application gateway for all communication, both inbound and outbound, is known as a trusted gateway, screened host firewall, or trusted application gateway. In a trusted gateway, certain applications are identified as trusted and are able to bypass the application gateway entirely and establish connections directly rather than by proxy. In this way, outside users can access information and Web servers without tying up the proxy applications on the application gateway.

Screened Subnet Firewall (DMZ)

Rather than using only the packet filtering router as the front door to the DMZ, a second firewall is added behind the packet filtering router to further inspect all traffic bound to or from the DMZ. The initial application gateway or firewall still protects the perimeter between the DMZ and the Intranet or secure internal network. The DMZ still contains mail, Web, and often e-commerce servers.

Multitiered DMZ

As e-commerce and e-business have proliferated, the need for e-commerce servers to access more and more secure information from database and transaction servers within the secure intranet has increased proportionately. As a result, the number of connections allowed through firewalls into the most secure areas of corporate networks has increased dramatically. In response to this phenomenon, the model of a multitiered DMZ has developed. Such a scenario really builds on the screened subnet firewall (DMZ) architecture by adding additional tiers to the DMZ, each protected from other tiers by additional firewalls. Typically, the first tier of the DMZ closest to the Internet would contain only the presentation or Web portion of the e-commerce application running on Web servers.

A firewall would separate that first tier of the DMZ from the second tier that would house the business logic and application portions of the e-commerce application typically running on transaction servers. This firewall would have rules defined to only allow packets from certain applications on certain servers with certain types of requests through from the first tier DMZ to the second-tier DMZ. The third (most secure) tier of the DMZ is similarly protected from the second tier by a separate firewall. The servers in the third-tier DMZ would be database servers and should contain only the data necessary to complete requested e-commerce transactions. Finally, as in the screened subnet firewall architecture, a firewall would separate the third-tier DMZ from the intranet, or most secure corporate network. Figure 2 illustrates some of the aforementioned enterprise firewall architectures.

FIREWALL FUNCTIONALITY AND TECHNOLOGY ANALYSIS

Commercially available firewalls usually employ either packet filtering or proxies as firewall architecture and add an easy-to-use graphical user interface (GUI) to ease the configuration and implementation tasks. Some firewalls even use industry standard Web browsers as their GUIs. Several certifying bodies are available to certify various aspects of firewall technology. As an example, one certifying body certifies the following:

- That firewalls meet the minimum requirements for reliable protection
- That firewalls perform as advertised
- That Internet applications perform as expected through the firewall

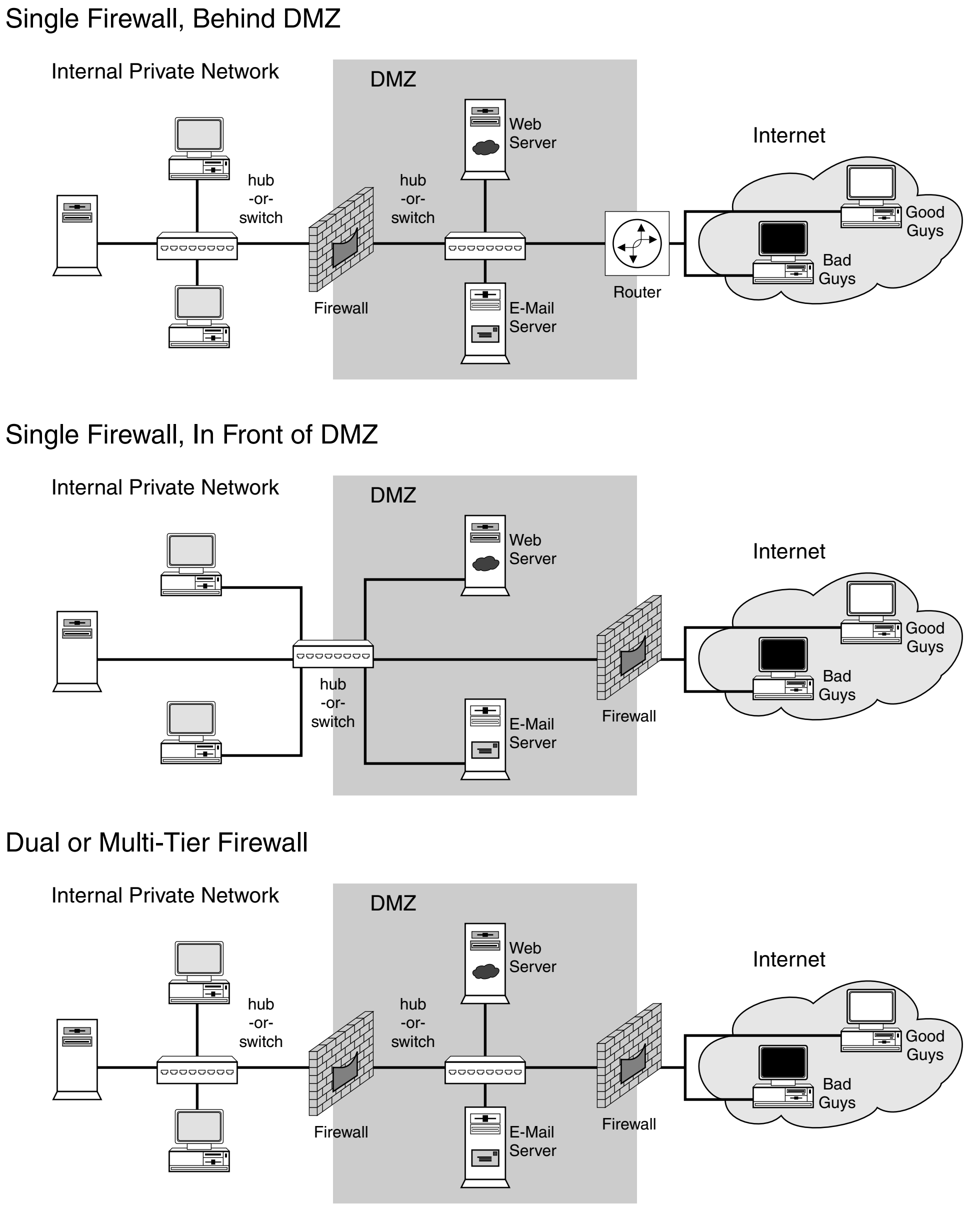


Figure 2: Enterprise firewall architectures.

Table 1 summarizes some of the key functional characteristics of firewall technology.

AIR GAP TECHNOLOGY

Some vendors try to assert the strength of their protection by using widely known terms such as "air gap" to characterize their protective mechanisms. One example of such a technology provides the same separation of the clean and dirty networks as firewalls. A hardware-based network switch is at the heart of this technology. The premise of creating a physical disconnection between the clean or secure network (Intranet) and the nonsecure or dirty network (Internet) is accomplished by connecting a server on the Internet connection that will receive all incoming requests. This server will connect to an electronic switch that strips the TCP headers and stores the packet in a memory bank, and then the switch disconnects from the external server and connects with the internal server. Once the connection is made, the internal server recreates the TCP header and transmits the packet to the intended server. Responses are made in the reverse order. The physical separation of the networks and the stripping of the TCP headers remove many of the vulnerabilities in the TCP connection-oriented protocol. Of course this provides little additional protection over other firewalls at the content level because content-level attacks are passed through the so-called air gap and responses are returned.

There is a protective device that is highly effective at limiting information flow to one direction. The so-called digital diode technology is applied in situations in which the sole requirement is that no information be permitted to leak from one area to the other while information

Table 1 Sample Functional Characteristics of Firewall Technology

FIREWALL FUNCTIONAL CHARACTERISTIC	EXPLANATION/IMPORTANCE
Encryption	Allows secure communication through firewall Encryption schemes supported: DES Encryption key length supported: 40, 56, 128 bits
Virtual Private Network (VPN) Support	Allows secure communication over the Internet in a virtual private network topology VPN Security protocols supported: IPsec
Application Proxies Supported	How many application proxies are supported? Internet application protocols (HTTP, SMTP, FTP, telnet, NNTP, WAIS, SNMP, rlogin, ping traceroute)? Real Audio? How many controls or commands are supported for each application?
Proxy Isolation	In some cases, proxies are executed in their own protected domains to prevent penetration of other proxies or the firewall operating system should a given proxy be breached
Operating Systems Supported	Unix and Varieties, Windows NT,2000,XP
Virus Scanning Included	Because many viruses enter through Internet connections, the firewall is a logical place to scan for viruses
Web Tracking	To ensure compliance with corporate policy regarding use of the World Wide Web, some firewalls provide Web tracking software. The placement of the software in the firewall makes sense because all Web access must pass through the firewall. Access to certain uniform resource locators (URLs) can be filtered
Violation Notification	How does the firewall react when access violations are detected? Options include SNMP traps, e-mail, pop-up windows, pagers, reports.
Authentication Supported	As a major network access point, the firewall must support popular authentication protocols and technology
Network Interfaces Supported	Which network interfaces and associated data link layer protocols are supported?
System Monitoring	Are graphical systems monitoring utilities available to display such statistics as disk usage or network activity by interface?
Auditing and Logging	Is auditing and logging supporting? How many different types of events can be logged? Are user-defined events supported? Can logged events be sent to SNMP managers?
Attack Protection	Following is a sample of the types of attacks that a firewall should be able to guard against: TCP denial-of-service attack, TCP sequence number prediction, source routing and RIP attacks, EGP infiltration and ICMP attacks, authentication server attacks, finger access, PCMAIL access, DNS access, FTP authentication attacks, anonymous FTP access, SNMP access remote access remote booting from outside networks; IP, MAC and ARP spoofing and broadcast storms; trivial FTP and filter to and from the firewall, reserved port attacks, TCP wrappers, gopher spoofing, and MIME spoofing
Attack Retaliation/Counterattack	Some firewalls can launch specific counterattacks or investigative actions if the firewall detects specific types of intrusions
Administration Interface	Is the administration interface graphical in nature? Is it forms based?

ARP = Address Resolution Protocol; DNS = Domain Name Server; EGP = Exterior Gateway Protocol; DES = Data Encryption Standard; FTP = File Transfer Protocol; HTTP = Hypertext Transfer Protocol; ICMP = Internet Control Message Protocol; IP = Internet Protocol; MAC = media access control; MIME = Multipurpose Internet Mail Extension; NNTP = Network News Transfer Protocol; RIP = Routing Information Protocol; SMTP = Simple Mail Transfer Protocol; SNMP = Simple Network Management Protocol; TCP = Transmission Control Protocol; WAIS = Wide Area Information Services.

is permitted to flow in the other direction. An example would be the requirement for weather information to be available to military planning computers without the military plans being leaked to the weather stations. This sort of protection typically uses a physical technology such as a fiber-optic device with only a transmitter on one end and receiver on the other end to provide high assurance of traffic directionality.

Finally, the real and widely accepted meaning of an "air gap" is a physical separation that is effective at eliminating all communications across the media.

SMALL OFFICE HOME OFFICE (SOHO) FIREWALLS

As telecommuting has boomed and independent consultants have set up shop in home offices, the need for firewalls for the SOHO market has grown as well. These devices are most often integrated with integrated services digital network–based multiprotocol routers that supply bandwidth on demand capabilities for Internet access. Some of these SOHO firewalls offer sophisticated features such as support for virtual private networks at a reasonable price. The most expensive of these costs less than $3,000 and some simpler filtering firewalls cost as little as $100. Some of these devices combine additional functionality such as network address translation, built in hub and switch ports, and load balancing in a combined hardware–software device known as a security appliance. Often DSL (digital subscriber line) or cable modems include firewall functionality because of their "always on" connection status.

Software-only solutions are also available. These products are installed on every computer and provide firewall-like functionality to each workstation. The benefits are that each workstation has an extremely inexpensive solution and that protection can be customized at the system level. The problem is that this requires all users to be firewall administrators for their computers, losing the economy of scale that was one of the original benefits associated with firewalls.

CONCLUSION

Firewalls are an essential and basic element of many organizations' security architecture. Nonetheless, technology must be chosen carefully to ensure that it offers the required functionality, and firewalls must be arranged into properly designed enterprise firewall architectures and properly configured, maintained, and monitored. Even in the best circumstances, firewalls should be seen as a relatively small solution within the overall information protection challenge and should not be considered sufficient on their own.

GLOSSARY

Address filtering A firewall's ability to block or allow packets based on internet protocol addresses.

Air gap technology Switched connections established to connect external and internal networks that are not otherwise physically connected.

Application-level gateway Application-level filters examine the entire request for data rather than only the source and destination addresses.

Bastion host Hardened server with trusted operating system that serves as the basis of the firewall.

Circuit-level proxies Proxy servers that work at the circuit level by proxying such protocols as file transfer protocol.

Demilitarized zone (DMZ) A neutral zone between firewalls or between a packet filtering router and a firewall in which mail and Web servers are often located.

Domain filtering A firewall's ability to block outbound access to restricted sites.

Dual-homed host A host firewall with two or more network interface cards with direct access to two or more networks.

Firewall A network device capable of filtering unwanted traffic from a connection between networks.

Internal firewall Firewalls that include filters that work on the data-link, network, and application layers to examine communications that occur only on an organization's internal network, inside the reach of traditional firewalls.

Multitiered DMZ A DMZ that segments access areas with multiple layers of firewalls.

Network address translation A firewall's ability to translate between private and public globally unique Internet Protocol addresses.

Packet filtering gateway A firewall that allows or blocks packet transmission based on source and destination Internet protocol addresses.

Port filtering A firewall's ability to block or allow packets based on transmission control protocol port number.

Proxy server Servers that break direct connections between clients and servers and offer application and circuit layer specific proxy services to inspect and control such communications.

Screened subnet Enterprise firewall architecture that creates a DMZ.

Stateful firewall A firewall that monitors connections and records packet information in state tables to make forwarding decisions in the context of previous transmitted packets over a given connection.

Trusted gateway In a trusted gateway, certain applications are identified as trusted, are able to bypass the application gateway entirely, and are able to establish connections directly rather than by proxy.

CROSS REFERENCES

See *Circuit, Message, and Packet Switching; Client/Server Computing; Guidelines for a Comprehensive Security System; Public Networks; Wide Area and Metropolitan Area Networks.*

FURTHER READING

Air gap technology. Whale Communications Web site. Retrieved from http://www.whalecommunications.com/fr_0300.htm

Canavan, John E. (2001). *Fundamentals of network security*. Boston: Artech House.

CERT Coordination Center. (1999, July 1). *Design the firewall system*. Retrieved from http://www.cert.org/security-improvement/practices/p053.html

Check Point Software Technologies. (1999, June 22). *Stateful inspection firewall technology tech note*. Retrieved from http://www.checkpoint.com/products/downloads/Stateful_Inspection.pdf

Edwards, J. (2001, May 1). Unplugging cybercrime. *CIO Magazine*. Retrieved from http://www.cio.com/archive/050100/development.html

Goldman, J. E., & Rawles, P. T. (2001). *Applied data communications: A business oriented approach* (3rd ed.). New York: Wiley.

Hurley, M. (2001, April 4). Network air gaps—drawbridge to the backend office. Retrieved from http://rr.sans.org/firewall/gaps.php

Keeping a safe distance. (2001, October). *Enterprise Technology*. Retrieved from http://www.avcom.com/et/online/2001/oct/keeping.html

NetGap. (n.d.) SpearHead Security Web site. Retrieved from http://www.spearheadsecurity.com/products.shtml

Scheer, Steven (2001, January 10). Israeli start-up may thwart Internet hackers. *InfoWorld*. Retrieved from http://staging.infoworld.com/articles/hn/xml/01/01/10/010110hnthwart.xml?Tem

Senner, L. (2001, May 9). Anatomy of a stateful firewall. *SANS Institute's Information Security Reading Room*. Retrieved from http://rr.sans.org/firewall/anatomy.php

Fuzzy Logic

Yan-Qing Zhang, *Georgia State University*

INTRODUCTION

To promote the use of fuzzy logic in the Internet, the 2001 BISC (Berkeley Initiative in Soft Computing, http://www-bisc.cs.berkeley.edu/) International Workshop on Fuzzy Logic and the Internet (FLINT2001) was held at the University of California, Berkeley (August 14–18, 2001; Nikravesh & Azvine, 2001), and a BISC Special Interest Group on Fuzzy Logic in the Internet was formed. Zadeh noted that "fuzzy logic may replace classical logic as what may be called the brainware of the Internet" (Zadeh, 2000). Clearly, the intelligent e-brainware based on fuzzy logic plays an important role in e-commerce applications on the Internet.

To increase the quality of intelligence of e-business on the Internet, computational web intelligence (CWI) based on computational intelligence (CI) and Web technology (WT) was proposed at the Special Session on Computational Web Intelligence at FUZZ-IEEE2002 of World Congress on Computational Intelligence (Zhang & Lin, 2002). CWI can be used to make intelligent e-business applications on the Internet and wireless networks. Fuzzy Web intelligence (FWI) is an important technology for intelligent fuzzy Internet and smart fuzzy e-commerce applications.

"In July 1964, Zadeh, a well-respected professor in the department of electrical engineering and computer science at the University of California, Berkeley, spent the evening by himself after a dinner with friends was canceled. It was then that the idea of grade of membership, which is the backbone of fuzzy set theory, occurred to him" (Yen & Langari, 1999, pp. 4–5). Zadeh published the seminal paper on fuzzy sets in 1965 (Zadeh, 1965). From 1965 to 1976, many researchers made strides in developing various fuzzy techniques such as fuzzy multistage decision making (Bellman & Zadeh, 1970), fuzzy measures (Sugeno, 1974), fuzzy automata (Mizumoto & Tanaka, 1976), fuzzy switching function (Kandel, 1973), and fuzzy optimization (Zimmermann, 1975). The first important fuzzy control system was designed by Assilian and Mamdani in 1974 (Yen & Langari, 1999). From 1977 to 1987, researchers and engineers continued to develop fuzzy techniques and use them for real applications. For example, Sendai city's subway system using fuzzy logic appeared in 1987. Since then, the fuzzy boom in Japan has been having an increasing impact on real applications of fuzzy logic theory. For example, various fuzzy products (e.g., fuzzy rice cookers, refrigerators, cameras, washing machines, etc.) on the market today. Fuzzy logic has applications in management, economics, and marketing (Zopounidis, Pardalos, & Baourakis, 2001). The fuzzy e-business boom on the Internet will likely continue into the future.

Crisp Sets and Fuzzy Sets

A crisp set is one with a binary characteristic function and thus has no uncertainty. A fuzzy set (Zadeh, 1965) is a one with a fuzzy characteristic function (or membership function), in which fuzziness, or uncertainty, exists. The key difference between a crisp set and a fuzzy set is the type of characteristic function. Assuming U to be a universe of discourse consisting of discrete objects or continuous space, a crisp set and a fuzzy set are defined as follows.

Definition 1

A crisp set A in U is defined as a set of ordered pairs:

$$A = \{(x, \phi_A(x)) \mid x \in U\},$$

where $\phi_A(x)$ is the binary characteristic function, $\phi_A(x) = 1$ if $x \in A$, and $\phi_A(x) = 0$ if $x \notin A$. $\phi_A(x) \in \{0, 1\}$.

Definition 2

A fuzzy set $\tilde{A}$ in U is defined as a set of ordered pairs:

$$\tilde{A} = \{(x, \mu_{\tilde{A}}(x)) \mid x \in U\},$$

where $\mu_{\tilde{A}}(x)$ is the fuzzy membership function that maps x to a membership degree between 0 and 1 ($\mu_{\tilde{A}}(x) \in [0, 1]$).

For example, an age space U is defined by $U = \{1, 2, 3, \ldots\ldots, 99, 100\}$. A crisp set for "young ages" is *Crisp-Young* = {(1, 1), (2, 1), (3, 1), . . . , (24, 1), (25, 1), (26, 0), (27, 0), . . . , (99, 0), (100, 0)}. We can see that there is a sharp jump from 25 years old to 26 years old, that is, 25 years old is 100% young, but 26 years old is 100% not young. In other words, if an age is less than or equal to 25, then it belongs to the crisp set *Crisp-Young*, otherwise it doesn't belong to *Crisp-Young*.

On the other hand, a fuzzy set for "young ages" is *Fuzzy-Young* = $\{(1, 1.0), (2, 1.0), \ldots, (19, 1.0), (20, 0.9), (21, 0.8), (22, 0.7), (23, 0.6), (24, 0.5), (25, 0.4), (26, 0.3), (27, 0.2), (28, 0.1), (29, 0), \ldots, (99, 0), (100, 0)\}$. We can see that there is a gradual decrease of the fuzzy membership value from 19 years old to 29 years old. In this case, 25 years old is 40% young, and 26 years old is 30% young.

Boolean Logic and Fuzzy Logic

Boolean logic (or binary logic) uses two logical values: true and false, or 1 and 0. Three basic Boolean logical operations are "AND," "OR," and "NOT," defined as follows.

Definition 3

The Boolean AND of two Boolean variables X and Y for $X, Y \in \{0, 1\}$ is denoted by $X \cap Y$. The four logical operations are $0 \cap 0 = 0$, $0 \cap 1 = 0$, $1 \cap 0 = 0$, and $1 \cap 1 = 1$.

Definition 4

The Boolean OR of two Boolean variables X and Y for $X, Y \in \{0, 1\}$ is denoted by $X \cup Y$. The four logical operations are $0 \cup 0 = 0$, $0 \cup 1 = 1$, $1 \cup 0 = 1$, and $1 \cup 1 = 1$.

Definition 5

The Boolean NOT of a Boolean variable X for $X \in \{0, 1\}$ is denoted by $\overline{X}$. The two logical operations are $\overline{0} = 1$ and $\overline{1} = 0$.

Fuzzy logic uses a logical value between 0 and 1. In other words, fuzzy logic uses not only extreme truth values (0: 100% false and 1: 100% true) but also partial truth values within 0 and 1. Three basic fuzzy logical operations are "fuzzy AND," "fuzzy OR," and "fuzzy NOT," defined as follows.

Definition 6

The fuzzy AND of two fuzzy variables x and y for $x, y \in [0, 1]$ is defined by $x \cap y = \text{t-norm}(x, y)$.

Definition 7

The fuzzy OR of two fuzzy variables x and y for $x, y \in [0, 1]$ is defined by $x \cup y = \text{t-conorm}(x, y)$.

Definition 8

The fuzzy NOT of a fuzzy variable x for $x \in [0, 1]$ is denoted by $\overline{x} = \text{negation}(x)$.

For example, three traditional fuzzy logical operations are $x \cap y = \min(x, y)$, $x \cup y = \max(x, y)$, and $\overline{x} = 1 - x$.

Various more complex crisp (fuzzy) logical expressions can be constructed by using crisp (fuzzy) NOT, crisp (fuzzy) AND, and crisp (fuzzy) OR.

Precise Rules and Fuzzy Rules

A precise rule based on Boolean logic has crisp variables with precise values. For example, "if temperature is higher than or equal to 100, then it is a fever." If a temperature is 100.001, then it's a fever (i.e., 100% true), but if a temperature is 99.999, then it's not a fever (i.e., 100% false). As this example indicates, the precise rule is not always meaningful.

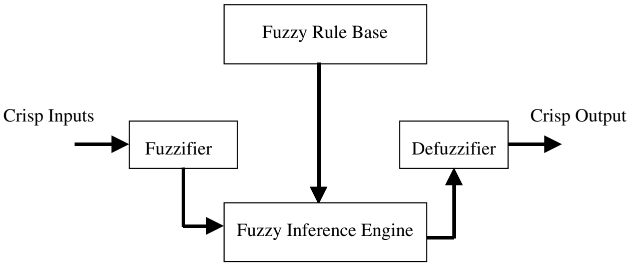


Figure 1: An example of fuzzy system architecture.

A fuzzy rule based on fuzzy logic has linguistic variables with fuzzy values. For example, "if temperature is *high*, then speed is *about 60 miles per hour*," and "if temperature is *low*, then speed is *about 30 miles per hour*." In this example, temperature and speed are two linguistic variables, temperature has two linguistic values *high* and *low*, and speed has two linguistic values *about 60 miles per hour* and *about 30 miles per hour*. These linguistic values are defined by fuzzy sets with relevant membership functions.

Based on the comparison between a precise rule and a fuzzy rule, we can see that a fuzzy rule is closer to a natural language and the natural knowledge in the human brain than a precise rule if we try to represent knowledge under uncertainty. In this sense, a fuzzy rule is more meaningful, more robust, and more useful than a precise rule.

Fuzzy Systems

A commonly used n-crisp-input-1-crisp-output fuzzy system (Figure 1) consists of four basic components: (a) a fuzzifier that can fuzzify inputs, (b) a fuzzy rule base that contains a set of fuzzy if–then rules, (c) a fuzzy inference engine that can generate fuzzy outputs for inputs by making fuzzy reasoning based on the fuzzy rule base, and (d) a defuzzifier that can defuzzify fuzzy conclusions into a crisp output.

There are different fuzzy systems such as Mamdani model (Jang, Sun, & Mizutani, 1997), TSK model (Yen & Langari, 1999), and Zhang–Kandel model (Zhang & Kandel, 1998). Because the Mamdani fuzzy system is a standard model that has been widely used in many applications, a simple 2-input–1-output Mamdani fuzzy system with a simple defuzzifier is discussed in the following section.

A fuzzy rule base contains 9 fuzzy if–then rules:

Rule 1: If x is low and y is low, then z is low,
Rule 2: If x is low and y is medium, then z is low,
Rule 3: If x is low and y is high, then z is medium,
Rule 4: If x is medium and y is low, then z is low,
Rule 5: If x is medium and y is medium, then z is medium,
Rule 6: If x is medium and y is high, then z is high,
Rule 7: If x is high and y is low, then z is medium,
Rule 8: If x is high and y is medium, then z is high,
Rule 9: If x is high and y is high, then z is high.

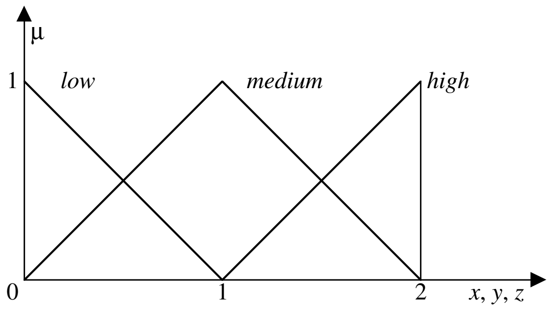


Figure 2: Three fuzzy linguistic values.

There are many types of fuzzy sets using different fuzzy membership functions, for example, trapezoidal fuzzy sets using trapezoidal fuzzy membership functions and Gaussian fuzzy sets using Gaussian fuzzy membership functions. For simplicity, three linguistic variables—x, y, and z—have the three same linguistic values *low*, *medium*, and *high*, defined by the three triangular fuzzy sets in Figure 2. The universe of discourse U is [0, 2], so x, y, $z \in [0, 2]$.

Let's assume $x = 0.25$ and $y = 1.5$ and see how a fuzzy system will generate a crisp output z. The procedure is given step by step as follows.

Step 1: Fuzzification

In Figure 3, the crisp input $x = 0.25$ is related to two linguistic values *low* with the degree 0.75 and *medium* with the degree 0.25, and the crisp input $y = 1.5$ is related to two linguistic values *medium* with the degree 0.5 and *high* with the degree 0.5.

Step 2: Fuzzy Inference

Because x is 75% low or 25% medium and y is 50% medium or 50% high, the four fuzzy Rules 2, 3, 5, and 6 are fired. The "if" part of our rule implies fuzzy AND of x and y, so $x \cap y = \min(x, y)$. In general, a t-norm operation such as $\min(x, y)$ used below can be used for the fuzzy AND.

Rule 2: If x is 0.75 low and y is 0.50 medium, then z is min(0.75, 0.50) = 0.50 low,

Rule 3: If x is 0.75 low and y is 0.50 high, then z is min(0.75, 0.50) = 0.50 medium,

Rule 5: If x is 0.25 medium and y is 0.50 medium, then z is min(0.25, 0.50) = 0.25 medium,

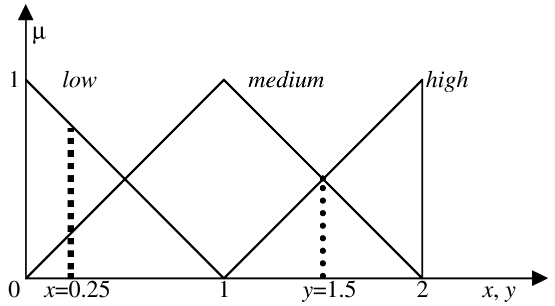


Figure 3: Fuzzifications for $x = 0.25$ and $y = 1.5$.

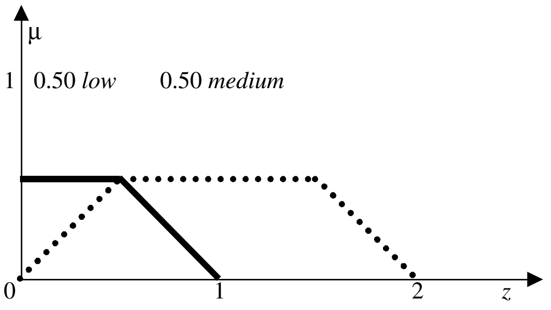


Figure 4: Fuzzy outputs of Rules 2 and 3.

Rule 6: If x is 0.25 medium and y is 0.50 high, then z is min(0.25, 0.50) = 0.25 high.

Graphical fuzzy inference results of Rules 2 and 3 are shown in Figure 4, and graphical fuzzy inference results of Rules 5 and 6 are shown in Figure 5. Therefore, four fuzzy outputs (fuzzy sets) are four partial fuzzy conclusions.

Step 3: Defuzzification

To convert the four fuzzy outputs in Figures 4 and 5 into a single crisp output, a defuzzification process is needed. Several defuzzifiers such as the center of gravity (also called center of area or centroid of area), mean of maximum, and so forth are commonly used (Yen & Langari, 1999). The center of gravity is introduced here.

The four partial fuzzy conclusions in Figures 4 and 5 are fuzzy ORed (Max operation is applied) to generate a single fuzzy conclusion (fuzzy set) $\tilde{C}$ shown in Figure 6. Then the final crisp output z_{output} (the center of gravity) is

$$\begin{aligned} z_{output} &= \frac{\int_0^2 \mu_C(z) z\, dz}{\int_0^2 \mu_C(z)\, dz} \\ &= \frac{\int_0^{1.5} 0.5z\, dz + \int_{1.5}^{1.75} (2-z) z\, dz + \int_{1.75}^{2} 0.25z\, dz}{\int_0^{1.5} 0.5\, dz + \int_{1.5}^{1.75} (2-z)\, dz + \int_{1.75}^{2} 0.25\, dz} \\ &= 0.917. \end{aligned}$$

In general, there are four types of fuzzy systems: (a) a crisp-input–crisp-output fuzzy system, (b) a crisp-input–fuzzy-output fuzzy system, (c) a fuzzy-input–crisp-output fuzzy system, and (d) a fuzzy-input–fuzzy-output fuzzy

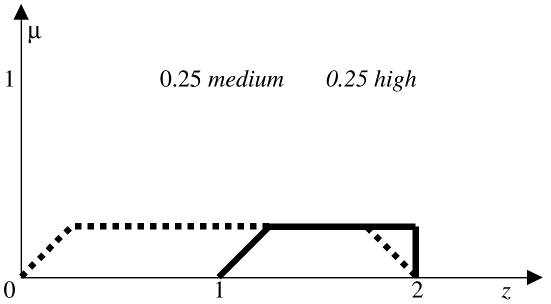


Figure 5: Fuzzy outputs of Rules 5 and 6.

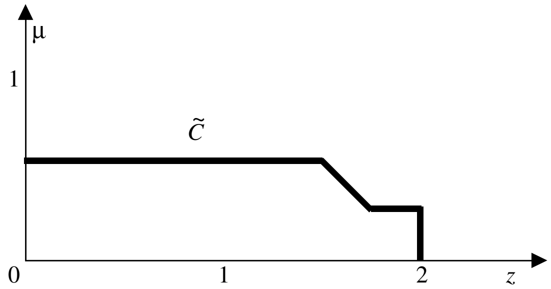


Figure 6: An integrated fuzzy output.

system. The previous example shows only how to design the crisp-input–crisp-output fuzzy system. Yen and Langari (1999) and Zhang and Kandel (1998) discuss the design of other fuzzy systems. The useful fuzzy logic Web sites are the BISC (http://www-bisc.cs.berkeley.edu/), International Fuzzy Systems Association (http://www.abo.fi/~rfuller/ifsa.html), and North American Fuzzy Information Processing Society (http://morden.csee.usf.edu/Nafipsf/).

FUZZINESS ON THE INTERNET

Data and information on the Internet are not always precise (100% accurate) but uncertain in many situations because of fuzziness, incompleteness, and other kinds of uncertainty. Fuzziness in e-commerce applications on the Internet is the major focus of this section.

E-commerce includes interactions at various levels between customers and business units, such as business to business, business to customers, and customers to customers. Communication and interactions of e-commerce are related to natural languages because customers use it as a basic tool. In general, natural languages have fuzziness in terms of words, syntax, semantics, and context. In addition, customers may have different interpretations of data and information because of fuzziness. For example, if two customers want to buy “very cheap” cars, a wealthy customer may consider a $30,000 car a “very cheap” one, whereas a less well-off customer may consider a $3,000 car “very cheap.” How to define “very cheap” is not a simple task because it is a fuzzy concept. Interestingly, fuzzy concepts are more useful and more robust than precise concepts in many e-commerce applications. For example, if a precise visitor searches accurate hotels using Google by typing in “$70” and “Hotel,” it is possible that no hotels with a rate of $70 per night will be found because the search criteria are *too* precise. If a fuzzy visitor types in “around $70” and “Hotel,” the fuzzy search engine may generate approximate robust search results such as rates of $70.05, $70.95, and so on. The visitor to the search engine may now choose the best of these fuzzy hotels.

> Fuzzy logic technology has the special ability to naturally represent human conceptualization and to make many useful contributions to the development of such a human-centered endeavors as e-commerce. Clearly, searching for information, services, and products will benefit from the facility of linguistic descriptions and partial matching available with fuzzy technologies. The business-to-business activity of automated procurement will benefit from the intelligent decision systems that can be constructed with fuzzy technology. (Yager, 2000)

FUZZY TECHNIQUES ON THE INTERNET

At the 2001 BISC International Workshop on Fuzzy Logic and the Internet (FLINT2001), Zadeh stated that fuzzy logic is the “brainware” of the Internet, and almost everything on the Internet is approximate in nature. Tim Barnes-Lee said, “Toleration of inconsistency can only be done by fuzzy systems. We need a semantic Web which will provide guarantees, and about which one can reason with logic. A fuzzy system might be good for finding a proof” (Zadeh, 2000). A Special Interest Group on Fuzzy Logic and the Internet was formed in 2001 (more detailed information can be found at http://www-bisc.cs.berkeley.edu/).

Almost any techniques can be fuzzified; for example, fuzzy expert systems, fuzzy reasoning, fuzzy optimization, fuzzy game theory, fuzzy pattern recognition. Many fuzzified techniques can be used on the Internet. Here, some sample applications are provided to show how fuzzy techniques can be used on the Internet.

Fuzzy Search Agents

Current search engines usually provide keyword-based searching tools using Boolean logic such as AND and OR. In most cases, several key words typed in by a user are ANDed for further searching. Such Boolean keyword-based search engines can return hundreds of pages, so it is very difficult to provide the user with most relevant Web links in the first several pages because precise keywords and Boolean operators have limited information for a search engine to find a small number of relevant Web links. In addition, some keywords contain uncertain information. For example, if a user types in “about $50” and “hotel”, a traditional search engine may use “about $50”, “hotel”, “about”, and “$50” to search relevant Web links, so returned Web links may not be satisfactory for the user because lots of hotels have “$49.95” per night instead of the precise “$50”. In this sense, the granular fuzzy Web search engine may return better meaningful search results using fuzzy words such as “about $50”, fuzzy operators such as fuzzy AND and fuzzy OR, and granular operators such as granular AND and granular OR (Zhang, Hang, Lin, & Yao, 2001).

Personalization is an important issue in the design of a search engine because different users may have different interests. For example, if different users input the same key word “Java”, a traditional search engine will return the same Web links, but a personalized search engine will return different Web links related to “Java” depending on ages of the users. Therefore, a personalized search engine containing a user’s profile can enhance quality of search results effectively. Furthermore, the personalized granular fuzzy Web search engine with dynamic Web mining

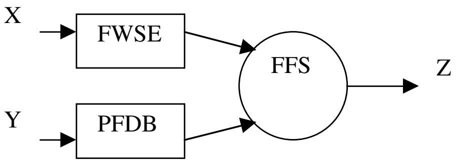


Figure 7: Logical architecture of the personalized fuzzy web search agent.

can better serve a user in terms of uncertainty and personalization.

Another problem is that many search tasks need context information to find more relevant and more meaningful search results. To deal with fuzzy context information, the conceptual fuzzy sets are introduced (Takagi & Tajima, 2001). The search engine based on conceptual fuzzy matching can search more relevant Web links (Takagi & Tajima, 2001). In general, a semantic search engine will play a more important role in Web searching than a linguistic search engine in the future.

The major challenging problem is that how to use fuzzy logic to design a fuzzy search agent using a basic search engine, users' profiles, and a fuzzy knowledge base to get more relevant and more meaningful search results. Here, a simple framework of the fuzzy search agent is introduced. Both the traditional keyword-based search method and the fuzzy personalized search algorithm are used to gain more satisfactory personalized search results for a particular user. In other words, different users may get different fuzzy search results even they type in the same key words. The basic architecture of the personalized fuzzy web search agent (PFWSA) is given in Figure 7. The PFWSA consists of the fuzzy Web search engine (FWSE) using fuzzy matching, the personalized fuzzy database (PFDB) including a user's profile, and the final fuzzy fusion system (FFS) making final selection and ranking using fuzzy reasoning. In Figure 7, X is a user's inputs (e.g., precise or fuzzy keywords and precise or fuzzy logical operators), Y is the updated user's personal information from either a user or a Web mining system, and Z is the final Web search results generated by the FFS based on results generated by the FWSE and personal information from the PFDB (Zhang, Hang, et al., 2001). In the future, how to actually implement the PFWSA will be a very interesting project. Importantly, the key thrust is how to design a better fuzzy user interface than a current precise user interface to make a user easily input fuzzy words and get more meaningful returned search results.

Here a simple fuzzy search algorithm is introduced to show how to design the key component of the PFWSA. To describe similarity between two fuzzy search key words like "around", "about", and "almost", a fuzzy relevancy matrix is created by using values between 0 and 1. To make a general purpose fuzzy relevancy matrix is very difficult, so here a small personalized fuzzy relevancy matrix is created for a single user. Initially, a personalized fuzzy relevancy matrix is created by experts and the user. Then the personalized fuzzy relevancy matrix is updated dynamically using the Web mining method.

The personalized fuzzy Web search algorithm is given as follows:

Step 1: Use fuzzy key words to find similar fuzzy key words based on the personalized fuzzy relevancy matrix, and then rank these similar fuzzy key words.

Step 2: Use these similar fuzzy key words to do regular Web search to find relevant results.

Step 3: Use the PFDB to select a small number of personalized results from the relevant results based on ranked personal preferences and context information if available.

Step 4: Return the final results in the order in terms of relevancy.

Fuzzy Information Agents

There are many types of Web information such as financial information, education information, shopping information, and science information. For particular Web information with fuzziness, a dedicated fuzzy information agent can serve an individual effectively based on the user's preferences. To show how a fuzzy Web information agent works, a personalized fuzzy stock information agent is introduced as follows.

Fuzzy logic implementation becomes complex when the number of inputs increases. Thus, to avoid system complexity and overhead while providing enough precision in output, a minimum number of inputs should be carefully selected. "Earning/share," "P/E Ratio," "dividend/share," and "yield" are chosen as inputs in the present system implementation among the values that can be fetched from Internet. The output of the fuzzy logic method is in the form of a list of predicted stock values. A general fuzzy rule has four input linguistic variables and one output linguistic variable. Each input linguistic variable has five linguistic values, so the total number of the fuzzy rule base is $5 \times 5 \times 5 \times 5 = 625$.

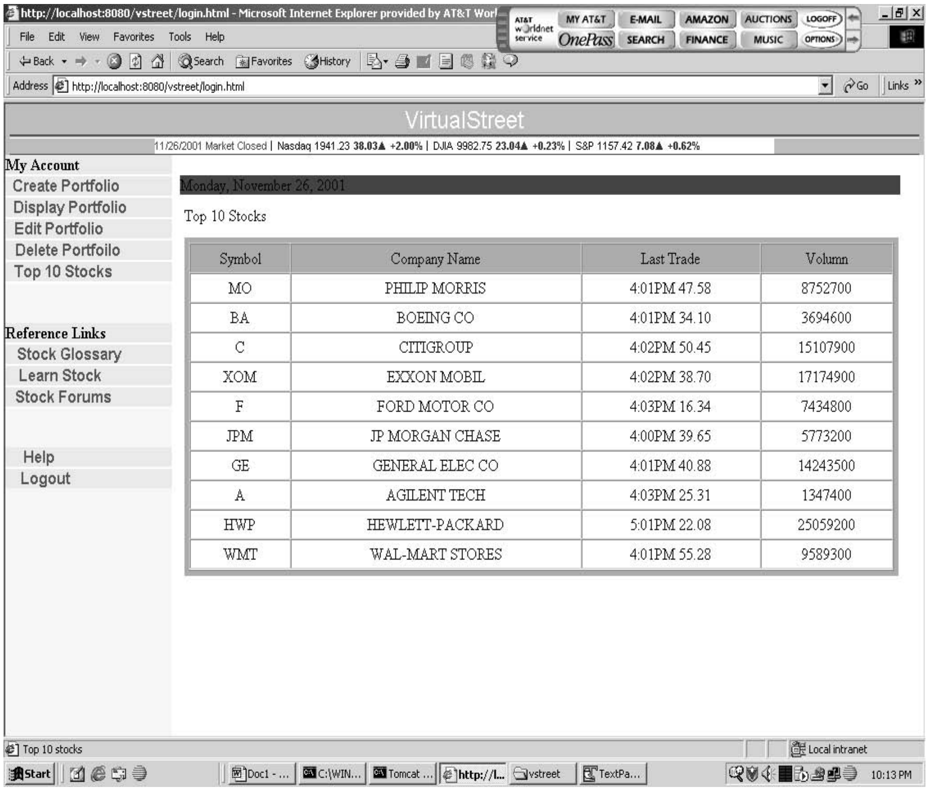
To get real-time stock information, a dial-up connection to the Internet is chosen to fetch stock information from Yahoo server at http://quote.yahoo.com. The trading data of each stock is fetched from Internet every 5 minutes. The services provided the personalized fuzzy stock information agent include portfolio management tools (create, display, edit, delete) and a list of "Top-10 Stocks" generated by a fuzzy inference system (Wang, Zhang, Belkasim, & Sunderraman, 2002).

Figure 8 shows the top-10 stocks determined by the fuzzy logic algorithm. The information in the list includes ticker symbol, company name, last trade time, price, and trade volume. The most attractive stock is listed first.

Fuzzy Intelligent Agents

The previous fuzzy stock information agent can provide a user with top-ranked stocks using fuzzy logic. Now a fuzzy intelligent agent can make a prediction and a decision. Following is an example of a fuzzy intelligent prediction agent.

A fuzzy neural Web-based stock prediction agent is developed using the granular neural network (GNN), which can discover fuzzy rules for future stock prediction (Zhang, Akkaladevi, Vachtsevanos, & Lin, 2001). To implement the fuzzy neural Web-based stock prediction agent, Java Servlets, Java Script, and JDBC (Java database connection) are used, and SQL (structured query language)


http://localhost:8080/vstreet/login.html - Microsoft Internet Explorer provided by AT&T Worl
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://localhost:8080/vstreet/login.html
Go Links
AT&T worldnet service
MY AT&T
E-MAIL
AMAZON
AUCTIONS
LOGOFF
OnePass
SEARCH
FINANCE
MUSIC
OPTIONS
Top 10 stocks
Local intranet
Start
Doc1 - ...
C:\WIN...
Tomcat ...
http://l...
vstreet
TextPa...
10:13 PM

VirtualStreet

11/26/2001 Market Closed | Nasdaq 1941.23 **38.03▲ +2.00%** | DJIA 9982.75 **23.04▲ +0.23%** | S&P 1157.42 **7.08▲ +0.62%**

My Account

Create Portfolio
Display Portfolio
Edit Portfolio
Delete Portfoilo
Top 10 Stocks

Reference Links

Stock Glossary
Learn Stock
Stock Forums

Help
Logout

Monday, November 26, 2001

Top 10 Stocks

Symbol	Company Name	Last Trade	Volumn
MO	PHILIP MORRIS	4:01PM 47.58	8752700
BA	BOEING CO	4:01PM 34.10	3694600
C	CITIGROUP	4:02PM 50.45	15107900
XOM	EXXON MOBIL	4:02PM 38.70	17174900
F	FORD MOTOR CO	4:03PM 16.34	7434800
JPM	JP MORGAN CHASE	4:00PM 39.65	5773200
GE	GENERAL ELEC CO	4:01PM 40.88	14243500
A	AGILENT TECH	4:03PM 25.31	1347400
HWP	HEWLETT-PACKARD	5:01PM 22.08	25059200
WMT	WAL-MART STORES	4:01PM 55.28	9589300

Figure 8: "Top-10 Stocks" ranked by fuzzy logic.

is used as the back-end database. The fuzzy neural Web-based stock prediction agent download real-time stock data from Yahoo, then converts the stock data into relational forms, and finally stores the relational stock data into a local database. Furthermore, the fuzzy neural Web-based stock prediction agent trains its internal GNN using the newly updated data, and then uses the trained GNN to predict future stock values. A Web user may choose different inputs to get different future predicted outputs.

The GNN can process various granules such as a class of numbers and an interval value. These granules (inputs and outputs of the GNN) are similar to multimedia data (inputs and outputs of biological neural networks in the human brain). The purpose of using the GNN is to simulate the functionality of the human brain so as to design an effective search agent.

The GNN has three inputs (open, high, and low values of past historical data of a company) and one output (the close amount for each day). A user may choose the number of input stock values (i.e., the beginning date and the last date of the historical stock data). After the GNN has been trained using the training data, the trained GNN can predict the future stock values.

For example, the GNN for stock prediction has three inputs x_1, x_2, and x_3, and one output y, where $y = f(x_1, x_2, x_3)$. x_1 is a fuzzy linguistic variable for open values, x_2 is a fuzzy linguistic variable for low values, x_3 is a fuzzy linguistic variable for high values, and y is a fuzzy linguistic variable for volume data. To make a Web user understand the predicted results easily, the graphical output user interface is designed to show the two-dimensional figures including both past data, trained data, and predicted data. An interesting future work is to embed the fuzzy neural Web-based stock prediction agent into a real-time online stock exchange system to make users use it to make decisions online. In general, an intelligent Web-based stock system can use fuzzy computing, granular computing, and neural computing to help users make reasonable decisions.

FUZZY E-COMMERCE

Fuzzy Web Shopping Agents

Various fuzzy Web shopping agents may have different functions and goals. A customer may use a fuzzy Web shopping agent to buy cheap or high-quality products, and a manager may use it to analyze customers' shopping data and then make smart decisions to make higher profits. Following is an example of a simple, manager-oriented fuzzy Web shopping agent using a fuzzy data mining method (Hearn & Zhang, 2001).

E-business has huge amounts of data for companies and costumers. To make more profit, a company needs to know costumer information by analyzing historical costumer data such as shopping habits and trends to make right business decisions. Without computers and Web mining tools, an e-business company could not find the useful costumer information quickly. So Web-based intelligent systems can greatly help employees analyze large amounts of e-business data efficiently. Unfortunately, the traditional Web-based intelligent system uses precise logic to do Web mining, so it could not process fuzzy data effectively. Because the human brain can process fuzzy data easily, an interesting problem is that how to use fuzzy logic to simulate the human brain in terms of fuzzy data mining. Therefore, analyzing the differences among fuzzy logic, traditional crisp logic, and human logic in e-business applications is an important for a design of a powerful e-business system. Here, some interesting experimental results are given for the comparison among fuzzy logic, traditional crisp logic and human logic. Based on the analysis, we will see which logic is suitable for e-business applications.

An online grocery store is used as a test bed. The data sets are generated from surveys and research. Four inputs are (1) amount of time spent online per week (in hours), (2) the number of online purchases made in the past six months, (3) the number of trips to the grocery store per week, and (4) amount spent (in dollars) at the grocery store per week. One output has five fuzzy categories (excellent, good, fair, poor, and very poor) to classify different users. So this is a 4-input–1-output classification system that can classify users into relevant fuzzy categories by using fuzzy logic, traditional crisp logic, and human logic, respectively.

The goal is to do effective marketing for relevant costumers. In this experiment, 25 "prospects" are created. Each prospect is classified by the three different types of logic. Table 1 shows the data. The three different logical systems are described below for comparison.

The first logical system is the fuzzy classification system that has four input fuzzy linguistic variables and one output fuzzy linguistic variable. An input fuzzy linguistic variable has three linguistic values (low, moderate, and high) and an output fuzzy linguistic variable has five linguistic values (excellent, good, fair, poor, and very poor). The fuzzy rule base has 81 TSK fuzzy rules (Yen and Langari, 1999).

The format of the first few TSK fuzzy rules is shown as follows:

If "time online" is low and "purchases" is low and "trips" is low and "groceries" is low, then z = (p * "time online input value") + (q * "purchases input value") + (r * "trips input value") + (s * "groceries input value") + t

If "time online" is low and "purchases" is low and "trips" is low and "groceries" is moderate, then z = (p * "time online input value") + (q * "purchases input value") + (r* "trips input value") + (s* "groceries input value") + t

If "time online" is low and "purchases" is low and "trips" is low and "groceries" is high, then z = (p * "time online input value") + (q * "purchases input value") + (r * "trips input value") + (s * "groceries input value") + t

If "time online" is low and "purchases" is low and "trips" is moderate and "groceries" is low, then z = (p * "time online input value") + (q * "purchases input value") + (r * "trips input value") + (s * "groceries input value") + t . . . ,

and so on.

In the above fuzzy rules, p, q, r, s, and t are initially set by experts. When an input data set is given, the data set

Table 1 The 25 Prospects for an Online Grocery Store

Customer	Time Online	Online Purchases	Store Trips	$ Spent at Store
1	4	8	2	$80
2	2	4	2	$125
3	9	2	0	$50
4	15	10	1	$100
5	0	0	3	$175
6	18	0	4	$150
7	5	0	2	$110
8	20	7	1	$50
9	2	0	3	$165
10	10	3	5	$180
11	1	1	7	$130
12	13	5	0	$25
13	8	9	2	$140
14	7	3	2	$100
15	3	0	3	$170
16	12	4	1	$80
17	20	10	1	$75
18	16	0	4	$145
19	9	2	2	$200
20	1	0	1	$185
21	14	1	0	$50
22	4	3	5	$200
23	6	7	3	$75
24	18	0	2	$40
25	2	5	6	$155

Table 2 Classification Rules

OVERALL Z VALUE	CLASSIFICATION
[0, 35)	Very poor
[35, 38)	Poor
[38, 42)	Fair
[42, 50)	Good
[50, 100)	Excellent

is mapped to a relevant category. The fuzzy classification system can automatically read input data from a data file, then use fuzzy reasoning to calculate the output value, and finally find out a suitable category based on the output value. For future use, the generated classification data sets are stored in an output file.

If a rule is fired, then the z value and w value for the rule must be calculated. The z value is determined based on the "then" part of the fuzzy rule. The w value is the minimum firing strength of the fuzzy sets in the rule. If the rule is not fired, the w and z values are set to 0.

If a fuzzy rule is fired for a new input data set, then the typical value z_i and the firing strength w_i for all rules are computed. The typical value z_i is determined based on the output fuzzy membership function. For example, the typical value z_i is given when the output fuzzy membership function $\mathrm{Fm}(z_i) = 1$. If fuzzy rule i is not fired, then the firing strength $w_i = 0$.

The final output Z for the prospect is calculated using the following defuzzification formula:

$$z = \frac{\sum_{i=1}^{81} w_i z_i}{\sum_{i=1}^{81} w_i}$$

The final classification is decided based on this overall Z using Table 2.

The second logical system is a crisp classification system using the same data sets for fair comparison. The crisp logic is used to make sharp classification for given input data directly based on Table 3.

The third logical system is the human classification system. The generated results by the human classification system are given in Table 4.

Table 3 Crisp Classification Rules

VARIABLE	RANGE	VALUE
Time online per week	0–5 hours	Low
	6–15 hours	Moderate
	16–20 hours	High
Online purchases in past six months	0–1 purchases	Low
	2–5 purchases	Moderate
	6–10 purchases	High
Trips to grocery store per week	0 trips	Low
	1–2 trips	Moderate
	3–7 trips	High
Amount spent in grocery store per week	$0–$75	Low
	$76–$140	Moderate
	$141–$200	High

Table 4 Comparisons Among Fuzzy Logic, Crisp Logic, and Human Logic

CUSTOMER	HUMAN	FUZZY	CRISP
1	Good	Good	Excellent
2	Good	Fair	Good
3	Poor	Very poor	Fair
4	Excellent	Excellent	Good
5	Very poor	Very poor	Poor
6	Good	Good	Excellent
7	Poor	Very poor	Very poor
8	Fair	Good	Very poor
9	Very poor	Poor	Poor
10	Good	Excellent	Excellent
11	Poor	Poor	Very poor
12	Very poor	Very poor	Fair
13	Excellent	Excellent	Excellent
14	Good	Fair	Excellent
15	Very poor	Poor	Poor
16	Fair	Poor	Good
17	Excellent	Excellent	Very poor
18	Good	Fair	Excellent
19	Excellent	Excellent	Excellent
20	Very poor	Poor	Poor
21	Poor	Very poor	Very poor
22	Excellent	Excellent	Excellent
23	Good	Fair	Fair
24	Poor	Very poor	Very poor
25	Excellent	Good	Excellent

To see the differences among the three logical classification systems, simulation results are summarized in Table 4. From Table 4, we can see that the fuzzy classification system is much closer to the human classification system than the crisp classification system. The crisp classification system may make some wrong decisions because the decision is too precise. The fuzzy classification system can make fault-tolerant decisions based on fuzzy rules as a human classification system can make robust decision based on internal uncertain rules in the human brain.

Fuzzy Mobile Wireless E-commerce

The wireless networks are growing rapidly. One may use a small handheld such as a personal digital assistant (PDA), Palm, or cell phone to do e-commerce such as mobile wireless shopping and mobile wireless stock exchanging. Such a small handheld device has limitations, however, such as low-speed processor, small memory, small communication bandwidth, and small screen. Fuzzy logic can be used to solve these problems by simplifying data classification, information processing, and decision making. There will be many application opportunities for fuzzy wireless e-commerce and fuzzy mobile e-commerce in the future. For example, the fuzzy reinforcement learning methods are used for power control in wireless transmitters (Vengerov & Berenji, 2002).

Fuzzy Game Theory in E-commerce

Classical game theory has been used in economy and business such as negotiation and auctions. Because data and information in e-commerce applications have much fuzziness, fuzzy game theory can play an important role in such fuzzy e-commerce applications. Fuzzy moves theory (Zhang et al., 1998), a novel fuzzy game theory, can be applied to e-commerce applications because there are many possible e-commerce games among customers, managers, human agents, and even software agents—for example, (a) fuzzy negotiation between customers and managers, between customers and customers, or between managers and managers; (b) fuzzy auction among customers; (c) fuzzy marketing games among different companies, and so on.

FUZZY WEB INTELLIGENCE IN E-COMMERCE

With explosive growth of the Internet, wireless networks, Web databases, and wireless mobile devices, Web users suffer from too many irrelevant Web search results (pages, links, etc.), fraud e-business transactions, nonpersonalized Web information, even wrong Web decisions. WI (Web intelligence) is studied carefully from different aspects and exploits artificial intelligence (AI) and advanced information technology on the Web and Internet (Yao, Zhong, Liu, & Ohsuga, 2001). Conventionally, AI is not the same as CI, although there is an overlap between them (Bezdek, 1998). CI has already been used in telecommunication applications (Pedrycz & Vasilakos, 2001).

CWI uses CI and Web technology (WT) to make intelligent e-Business applications on the Internet and in wireless networks. Specifically, the key focus of CWI is how to handle the uncertainty of smart e-business systems effectively to make the Internet and wireless networks intelligent.

Computational Web Intelligence

In basic terms, CWI is a hybrid technology of CI and WT that is dedicated to increasing the quality of intelligence of e-business applications on the Internet and in wireless networks. Currently, the seven major research areas of CWI are (a) fuzzy WI, (b) neural WI, (c) evolutionary WI, (d) probabilistic WI, (e) granular WI, (f) rough WI, and (g) hybrid WI. In the future, more CWI research areas will no doubt evolve.

Fuzzy Web Intelligence

FWI comprises two major techniques: (a) fuzzy logic and (b) WT. The main goal of FWI is to design intelligent fuzzy e-agents that can deal with the fuzziness of data, information, and knowledge and make satisfactory decisions for e-business applications similar to those the human brain would make. There are various applications of FWI. For example, fuzzy intelligent agents are used to help advertisers make better marketing decisions for Web site visitors. Effective FWI applications for the Internet and e-commerce are current goals. Just as rule-based models are central to the fuzzy logic used within the Web environment, so, too, is optimization of fuzzy Web systems an important goal. General optimization tools, including other techniques such as neural networks and genetic algorithms can enhance the performance of the fuzzy Web systems.

CONCLUSION

With explosive growth of e-commerce on the Internet and wireless networks, both users and managers suffer from novel wired and wireless application problems in terms of quality of intelligence and service. To improve these aspects of e-commerce and e-business, the intelligent fuzzy e-brainware using fuzzy logic techniques should be researched and developed.

Computational web intelligence—merging CI and WT—will play an important role in future intelligent e-business applications on the Internet and in wireless networks. FWI is the fundamental technology for building smart, fuzzy wired and wireless e-commerce applications. The goal of using fuzzy logic and other granular and soft computing techniques is to build intelligent Internet and smart e-commerce systems on wired and wireless networks to help diverse users conduct intelligent e-business effectively, efficiently, and safely.

GLOSSARY

Computational intelligence (CI) Broad intelligent technology including fuzzy logic, neural networks, evolutionary algorithms, probabilistic methods, granular computing, and rough set theory; soft computing is the core of CI (Yager & Zadeh, 1994).

Computational Web intelligence (CWI) A hybrid technology of computational intelligence and Web technology dedicating to increasing quality of intelligence of e-business applications on the Internet and in wireless networks.

Fuzzy logic Nontraditional logic using fuzzy variables with a truth value between 0 and 1; the three basic fuzzy logical operations are fuzzy AND (t-norm), fuzzy OR (t-conorm), and fuzzy NOT (negation).

Fuzzy reasoning The notion that a method can logically generate fuzzy conclusions from given crisp or fuzzy inputs based on a fuzzy rule base.

Fuzzy sets A generalized set characterized by a mapping from its universe of discourse into the interval [0, 1]; the degree of how an element belongs to a fuzzy set is represented by a membership function (This is the most common notion of fuzzy sets; others include type II fuzzy sets [Zadeh, 1973a, 1973b, 1973c], qualitative fuzzy sets [Lin, 2001; Thiele, 1999], and granular fuzzy sets (probability-based fuzzy sets) [Lin, 1989, 1997]).

Fuzzy Web intelligence (FWI) A fuzzy Web technology dedicated to increasing the quality of intelligence of e-business applications on the Internet and in wireless networks.

CROSS REFERENCES

See *Intelligent Agents; Mobile Commerce; Rule-Based and Expert Systems.*

REFERENCES

Bellman, R. E., & Zadeh, L. A. (1970). Decision making in a fuzzy environment. *Management Science, 17*, 141–164.

Berkeley Initiative in Soft Computing Web site (n.d.). Retrieved August 18, 2002, from http://www-bisc.cs.berkeley.edu/

Bezdek, J. C. (1998). Computational intelligence defined—by everyone! In O. Kaynak, L. A. Zadeh, B. Turksen, & I. J. Rudas, (Eds.), *Computational intelligence: Soft computing and fuzzy-neuro integration with applications* (pp. 10–37). Berlin, Germany: Springer.

Hearn, K. L., & Zhang, Y.-Q. (2001). Fuzzy, crisp, and human logic in e-commerce marketing data mining. *Proceedings of The International Society for Optical Engineering (SPIE) AeroSense 2001: Conference of Data Mining and Knowledge Discovery: Theory, Tools, and Technology* (Vol. 4384, pp. 67–74). Bellinham, Washington: SPIE.

International Fuzzy Systems Association Web site (n.d.). Retrieved August 18, 2002, from http://morden.csee.usf.edu/Nafipsf/http://www.abo.fi/~rfuller/ifsa.html

Jang, J.-S. R., Sun, C.-T., & Mizutani, E. (1997). *Neuro-fuzzy and soft computing: A computational approach to learning and machine*. Upper Saddle River, NJ: Prentice-Hall.

Kandel, A. (1973). On minimization of fuzzy functions. *IEEE Transactions on Computers, C-22*, 826–832.

Lin, T. Y. (1989). Neighborhood systems and approximation in database and knowledge base systems. *Proceedings of the Fourth International Symposium on Methodologies of Intelligent Systems* (pp. 75–86). Amsterdam, The Netherlands: North Holland.

Lin, T. Y. (1997). Neighborhood systems—a qualitative theory for fuzzy and rough sets. In P. Wang (Ed.), *Advances in machine intelligence and soft computing* (Vol. IV, 132–155). Raleigh, NC: Duke University.

Lin, T. Y. (2001). Qualitative fuzzy sets: A comparison of three approaches. *Proceedings of the Joint 9th* International Fuzzy Systems Association *World Congress and 20th* North American Fuzzy Information Processing Society *International Conference* (pp. 2359–2363). Piscataway, NJ: Institute of Electrical and Electronics Engineers (IEEE).

Mizumoto, M., & Tanaka, K. (1976). Fuzzy-fuzzy automata. *Kebernets, 5*, 107–112.

Nikravesh, M., & Azvine, B. (2001). New directions in enhancing the power of the Internet. *Proceedings of the 2001 Berkeley Initiative in Soft Computing International Workshop on Fuzzy Logic and the Internet*. Berkeley, CA: Univ. of California.

North American Fuzzy Information Processing Society Web site (n.d.). Retrieved August 18, 2002, from http://morden.csee.usf.edu/Nafipsf/

Pedrycz, W., & Vasilakos, A. (2001). *Computational intelligence in telecommunications networks*. Boca Raton, FL: CRC Press.

Sugeno, M. (1974). *Theory of fuzzy integrals and its applications*. Unpublished doctoral dissertation, Tokyo Institute of Technology.

Takagi, T., & Tajima, M. (2001). Proposal of a search engine based on conceptual matching of text notes. *Proceedings of the 2001 Berkeley Initiative in Soft Computing International Workshop on Fuzzy Logic and the Internet* (pp. 53–58). Berkeley, CA: Univ. of California.

Thiele, H. (1999). On the concepts of the qualitative fuzzy sets. *Proceedings of 1999 IEEE International Symposium on Multiple-Valued Logic* (pp. 282–287). IEEE Computer Society.

Vengerov, D., & Berenji, H. R. (2002). Using fuzzy reinforcement learning for power control in wireless transmitters. *Proceedings of 2002 IEEE International Conference on Fuzzy Systems of World Congress on Computational Intelligence 2002* (pp. 797–802). New York: IEEE.

Wang, Y., Zhang, Y.-Q., Belkasim, S., & Sunderraman, R. (2002). Real time fuzzy personalized Web stock information agent. *Proceedings of the 2nd International Workshop on Intelligent Systems Design and Applications* (pp. 83–87). Atlanta, GA: Dynamic Publishers, Inc.

Yager, R. R. (2000, November/December). Targeted e-commerce marketing using fuzzy intelligent agents. *IEEE Intelligent Systems*, 42–45.

Yager, R. R., & Zadeh, L. A. (1994). *Fuzzy sets, neural networks, and soft computing*. New York: Van Nostrand Reinhold.

Yao, Y. Y., Zhong, N., Liu, J., & Ohsuga, S. (2001). Web intelligence (WI): Research challenges and trends in the new information age. In Zhong, N., Yao, Y. Y., Liu, J., & Ohsuga, S. (Eds.), *Web Intelligence: Research and Development. Proceedings of the First Asia-Pacific Conference, WI 2001* (pp. 1–17). Berlin, Germany: Springer.

Yen, J., & Langari, R. (1999). *Fuzzy logic: Intelligence, control, and information*. Upper Saddle River, NJ: Prentice-Hall.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control, 8*, 338–353.

Zadeh, L. A. (1973a). The concept of a linguistic variable and its application to approximate reasoning—I. *Information Sciences, 8*, 199–249.

Zadeh, L. A. (1973b). The concept of a linguistic variable and its application to approximate reasoning—II. *Information Sciences, 8*, 301–357.

Zadeh, L. A. (1973c). The concept of a linguistic variable and its application to approximate reasoning—III. *Information Sciences, 9*, 43–80.

Zadeh, L. A. (2000). Welcome to The BISC Special Interest Group on Fuzzy Logic and the Internet. Retrieved from http://www.cs.berkeley.edu/~nikraves/bisc/sig/internet/msglaz2.htm

Zhang, Y.-Q., Akkaladevi, S., Vachtsevanos, G., & Lin, T. Y. (2001). Fuzzy neural web agents for stock prediction. *Proceedings of the 2001 Berkeley Initiative in Soft Computing International Workshop on Fuzzy Logic and the Internet* (pp. 101–105). Berkeley, CA: Univ. of California.

Zhang, Y.-Q., Hang, S., Lin, T. Y., & Yao, Y. Y. (2001). Granular fuzzy web search agents. *Proceedings of the 2001 Berkeley Initiative in Soft Computing International Workshop on Fuzzy Logic and the Internet* (pp. 95–100). Berkeley, CA: Univ. of California.

Zhang, Y.-Q., & Kandel, A. (1998). *Compensatory genetic fuzzy neural networks and their applications* (Series in Machine Perception Artificial Intelligence, Volume 30). Singapore: World Scientific.

Zhang, Y.-Q., & Lin, T. Y. (2002). Computational web intelligence (CWI): Synergy of computational intelligence and web technology. *Proceedings of 2002 IEEE International Conference on Fuzzy Systems* (pp. 1104–1107). New York: IEEE.

Zimmermann, H. J. (1975). Description and optimization of fuzzy systems. *International Journal of General Systems, 2*, 209–215.

Zopounidis, C., Pardalos, P. M., & Baourakis, G. (2001). *Fuzzy sets in management, economics and marketing*. Singapore: World Scientific.