INTERNET TECHNOLOGY

There are two distinct ways to define the World Wide Web (web). One way is by the manner in which it creates a unified electronic publishing platform. Hypertext markup language (HTML) is a tagging convention for displaying information contained in a specially encoded text document. The basic document of the web is called a page. While the web is primarily a hypermedia publishing platform, additional functionality can be achieved through the use of such elements as common gateway interfaces (CGI), Java scripting, and add-on software applications.

Through a uniform resource locator (URL), which is contained the markup instruction, a web browser is able to locate a designated resource. The instructions contained in a web page can include hypertext (hyperlink) pointers to other web pages. Basic HTML elements include placement of graphics and text, color attributes (font and background), and designated fonts. Other HTML elements can activate applets (small applications that are added into the browser, commonly known as plug-ins or ActiveX controls) or external applications (also known as helper applications) such as word processors, presentation programs, or other programs independent of the browser.

As a publishing platform the web remains without parallel in the traditional forms of media. The web is considered a 7-day-a-week, 24-hour-a-day international publishing environment. The web is also the most egalitarian of publishing forums. Web sites (a collection of web pages) can be run by multibillion-dollar corporations or by individuals. Because the web permits multimedia, including sound, video, virtual reality, and interactive programming, all forms of traditional media are represented on the web.

Another way to define the web is by its basic technical structure. Technically the web uses a data access protocol (also called request/response or client/server), called Hypertext Transfer Protocol (HTTP). This protocol was designed specifically for the efficient distribution of hypertext. HTTP can be used on the Internet or an intranet. The Internet is a worldwide collection of computer networks that uses the Transmission Control Protocol/Internet Protocol. An intranet is a web server that is blocked by a firewall to the Internet.

A web browser (sometimes called a "user agent") is the client side of the web. The browser uses the HTTP to request documents from the server. While many browsers were developed during the short history of the web, today the two dominant browsers are Netscape Navigator (Netscape Communication Corporation) and Internet Explorer (Microsoft Corporation). The browser is able to interpret the highest version of the various standards that was embedded into its program when the browser's code was written. As this article was written the current version for both browsers is 4.x.

EARLY HISTORY OF THE WORLD WIDE WEB

The World Wide Web, much like the Internet itself, was more a product of evolution than one of outright planning. In 1980, while a consultant for CERN, the European Laboratory for Particle Physics in Geneva, Switzerland. Tim Berners-Lee wrote a notebook program, "Enquire-Within-Upon-Everything," allowing links to be made between arbitrary nodes. Each node had a title, a type, and a list of bidirectional typed links. In 1989 Berners-Lee circulated "Information Management: A Proposal" for comments at CERN. With Robert Cailliau as co-author, the revised proposal, "World Wide Web: Proposal for a Hypertext Project," was presented in November 1990 to CERN. It was at this time that the name, World Wide Web, was born. Berners-Lee used the World Wide Web as a name for the project and the name stuck. It was during this period that Berners-Lee introduced the URL, HTTP, and HTML standards with prototype Unix-based servers and browsers.

Technical Student Nicola Pellow developed a line-mode browser that was released to a limited audience on "priam" vax, rs6000, and sun4 in 1991. General release of the web was released on the central CERN machines in May 1991. By 1993 Midas (Tony Johnson, SLAC), Erwise (HUT), and Viola

(Pei Wei, O'Reilly Associates) browsers are available for X; CERN Mac browser (ECP) released as alpha. In early 1993 there were approximately 50 known HTTP servers.

In February 1993, the NCSA at the University of Illinois released first alpha version of Marc Andreessen's "Mosaic for X." Mosaic was significant because it was the first browser to use a graphical interface. By September, World Wide Web traffic (Port 80 HTTP) measures 1% of the National Science Foundation's backbone traffic. In the same month, NCSA released working versions of the Mosaic browser for all common platforms: X, PC/Windows, and Macintosh.

During 1994 Marc Andreessen and colleagues had left the NCSA and formed Mosaic Communications Corp. which would eventually become Netscape Communications. In October the Massachusetts Institute of Technology and CERN agreed to start the World-Wide Web Consortium (W3C). The W3C was founded to develop common protocols for the web. However, in December the CERN Council approved the construction of the Large Hadron Collider accelerator. The commitment to accelerator imposes financial limitations and CERN decides not to continue development in the web.

THE WORLD WIDE WEB CONSORTIUM

The W3C is an international industry consortium with over 200 members. The organization is jointly hosted by the Massachusetts Institute of Technology Laboratory for Computer Science (MIT/LCS) in the United States; the Institut National de Recherche en Informatique et en Automatique (INRIA) in Europe; and the Keio University Shonan Fujisawa Campus in Asia. Tim Berners-Lee serves as the Director of the W3C and creator of the World Wide Web, and Jean-François is Chairman of the W3C.

The purpose of consortium is to find common standards and specifications for the web. Although principally known as the organization which develops the HTML recommendation, the W3 is involved in other areas of web development including the Platform for Privacy Preferences Project and Digital Signature Initiative. The W3C divides its development activities into three domains: user interface, technology and society, and architecture.

Specifications are developed within the W3C and is reviewed through the stages of Working Draft, Proposed Recommendation, and Recommendation. The documents are available at the W3C web site. The specifications must be formally approved by the membership.

HYPERTEXT MARKUP LANGUAGE

HTML is used to prepare hypertext documents to be distributed on the web. The web browsers interpret the HTML information and present it to the user. The recommendation for HTML is established by the World Wide Web Consortium. The protocol is nonproprietary, and the tag convention is based upon standard generalized markup language (SGML). SGML is an ISO standard (ISO 8879:1986) which supplies a formal notation for the definition of generalized markup languages. A simplified form of SGML, called XML (extensible markup language), which is optimized for the web, is under development.

All HTML is written in the American Standard Code for Information Interexchange (ASCII). HTML creation does not require the use of specific software; however, most authors use an HTML editor. HTML can also be created from many major software applications (such as Microsoft Word) which contain HTML converters. HTML filenames typically end with the extension .html or .htm. These identifers signal to the browser that the file is an HTML document.

HTML uses tags to define elements on a web page. The elements specify the meaning associated with a block of text or attributes. An attribute is a specifically defined property such as an image. The following HTML statement

⟨U⟩engineering⟨/U⟩

means underline the word engineering.

HTML elements can also take on attributes which usually have assigned meaning. An image element (IMG element), for example, inserts an image into the text but does change a block of text. The HTML statement ⟨IMG SRC = sample.gif⟩ would create a URL pointer to the image file, which in this example is called sample.gif. There have been two image types that the web browsers have understood: .gif (graphic interexchange format) and .jpeg (journalist photographic exchange graphic). The newest recommendation for HTML seeks to replace .gif with .sng (simple network graphics). File extensions play an important role in web publishing as the extensions inform the browser to perform certain functions, such as displaying an image or starting an application.

Version 3.0 of Microsoft Internet Explorer and Netscape Navigator added support for a ⟨SCRIPT⟩ tag which refers to client-side scripting. This allows web pages to contain small programs (such as Javascript) that provide gateways between the user and web server.

HYPERTEXT TRANSFER PROTOCOL

The Hypertext Transfer Protocol is a generic, application-level, object-oriented protocol designed for distributed information systems. HTTP is also used as a generic protocol for communication between user agents and proxies/gateways to other Internet protocols, which permits access to other Internet resources (such as mail and ftp) through a common interface.

HTTP connections have four stages. First the connection is opened when a user contacts the server with a URL. The browser then sends an HTTP request header to the server. The server then sends a HTTP response header which discusses the status of the response, and then the actual data are sent. The connection is then closed.

If a user requests a file, the HTTP server locates the files and sends it. If the user was to send information back to the server—for example, by filling in a form on a web page—the HTTP server passes this request to gateway programs. The specification for HTTP servers is called the common gateway interface (CGI). CGI permits server-side applications to be invoked and are referenced through URLs contained in a web page. CGI programs can be compiled programs or they can be executable scripts.

EXTENDING BASIC WORLD WIDE WEB FUNCTIONS

The functionality of web browsers can be extended by additional programs that are invoked by when specific file extensions are interpreted. A wide range of applications are included: audio, video, virtual reality, graphic viewers, animated graphics, and others. These additional applications are described in a number of ways: plug-ins, helper applications, apples, and ActiveX controls.

One of the most popular are Java applets. Java is high-level programming language developed by Sun Microsystems to be an object-oriented, architectural neutral way to distribute software. On the web, Java applets run within a Java-enabled web browser. Typically the application on the web is limited to added multimedia functionality to the web browser. While Java is more powerful than its simple use on the web suggests, its functionality on the web is not as sophisticated as platform-specific applications have proven to be.

The two largest arenas of applets are ActiveX controls and plug-ins. ActiveX controls are software components that download automatically when used on a web page. Plug-ins are downloaded and installed separately, and then the functionality is incorporated into a browser.

UNIFORM RESOURCE LOCATORS

Uniform resource locators (URLs) is the addressing scheme of the web. However, the URL scheme can be used for other protocols as well [such as FTP (file transfer protocol) and gopher]. URLs use a single line of ASCII characters. The URL has three main parts: The protocol specifier, the Internet domain name, and a path and file name to the document, although the latter part may not be needed.

PUSH TECHNOLOGY: HOW THE WEB IS USED

The web is home to many forms of information and communication exchange. While the number of web sites located outside of the United States is growing, web servers are still predominantly located in the United States or are owned by US companies. Perhaps the best way to define the functions of the web is to analyze the principal purpose of web sites, even though a single web site may share several purposes. A single web server, Shockrave for example, currently distributes interactive games, music, and animated cartoons.

The web is, first and foremost, a unified information service. Technically, all information on the web may be considered published information, even though it does not come from a traditional publisher or news organization. Thus, the only effective way to define the web's purposes is to examine the purpose of the information that is being distributed.

Search Engines and Directory Services

There are two approaches for finding information on the web: through the use of a search engine or through a directory. All of the search engines do keyword searching against a database, but results differ because of the manner in which the information is compiled. There are hundreds of search engines in a variety of languages available on the web. Search engines use web software agents (known as spiders or robots or crawlers) to automatically gather information from web sites. The agent identifies a page on a server, reads it, and also follows links to other pages. The agents return on a regular basis to update its entries.

The information found by the agent is collected in an index, also referred to as a catalog. The search engine software then sifts through the information to find matches and to rank relevancy. Because the web lacks a common controlled vocabulary and relies on information provided by the page creators, searching the web can be problematic. Most search engines on the web rely on relevance retrieval, a statistical means of identifying a page's relevance by the number of times a word appears in relationship to the number of words in a document. Word placement can also be a factor incorporated in the search algorithm.

A few of these search engines are dominant in popularity. Hotbot and Altavista are two of the largest search engines. All three of these search engines offer simple and advanced searching modes. The advanced searching mode includes Boolean operators and field-limiting capabilities.

The web directories are created by human beings and rely either on submissions or on site selection to create the database. While these directories typically have a search engine attached to them, and often include a statement "search the web," the database lookup is limited to the information contained in the database, not the entirety of the web. Increasingly, these sites are also offering free electronic mail in an effort to attract more visitors. Yahoo is the oldest of these services, begun in April 1994 by David Filo and Jerry Yang, then PhD candidates at Stanford University. Infoseek, Excite, Webcrawler, and Lycos offer similar services.

Other search engines on the web include multisearch databases which search for more than one database at a time. The web is also witnessing the growth of specialized directories:

Government Publishing
Library and Database Services
Educational Uses
Community Servers
Traditional Publishing
Scholarly Publishing
Electronic Commerce
Software Distribution
Technical Support
Interactive Chat
Interactive Gaming
Telephony

BIBLIOGRAPHY

All of the following resources are available online.

http://browserwatch.internet.com/

AltaVista (advanced) http://altavista.digital.com/

http://www.altavista.digital.com/av/content/about_our_story.htm

Infoseek Ultrasmart http://www.infoseek.com/

AltaVista (advanced) http://altavista.digital.com/cgi-bin/query?pg=aq&what=web

OpenText http://index.opentext.net/

Excite Search http://www.excite.com

HotBot http://www.hotbot.com/

Webcrawler http://www.webcrawler.com

Lycos http://www.lycos.com

Meta and Multi-Search Engines

Savvy Search Multi-Search http://guaraldi.cs.colostate.edu:2000/

Savvy Search search form http://guaraldi.cs.colostate.edu:2000/form

Metacrawler Multisearch

DogPile http:/www.dogpile.com

Inference Find http://www.inference.com/ifind/

Profusion MetaSearch http://www.designlab.ukans.edu/profusion/

Highway 61 Multisearch http://www.highway61.com

Beaucoup 600 Search Engines

Mamma Mother of All Search Engines http://www.mamma.com/

Cosmic Mother Load Insane Search http://www.cosmix.com/motherload/insane/

WebSearch MetaSearch http://www.web-search.com:80/

CNETs Search.com Multi-Search Page http://www.search.com

Webreference Search Engine page http://www.webreference.com/search.html

Specialized Search Engines

AT1 Database search: The invisible webhttp://www.at1.com/

EDirectory search engines from around the world http://www.edirectory.com/

Muscat EuroFerret European Site Search http://www.muscat.co.uk/euroferret/

International Regional Search Engines http://searchenginewatch.com/regional/

Search Net Happenings http://www.mid.net:80/NET/

Inquiry Com Information Technology search http://www.inquiry.com/

Mediafinder http://www.mediafinder.com/custom.cfm

Internic's Whois Domain Information http://ds.internic.net/wp/whois.html

Domain Name Search http://www.ibc.wustl.edu/ibc/domain_form.html

Study Web Research Site http://www.studyweb.com/

Library of Congress Search http://lcweb.loc.gov/harvest/

FindLaw Legal Search http://www.findlaw.com/index.html

Legal Search Engines http://www.uklaw.net/lawsearch.htm

InfoMine Government info search http://lib-www.ucr.edu/search/ucr_govsearch.html

HealthGate Free Medline http://www.healthgate.com/HealthGate/MEDLINE/search.shtml

Medical Matrix Medline Search http://www.medmatrix.org/info/medlinetable.html

Four11 People http:/www.four11.com

Forum One Forums http://www.forumone.com

DejaNews Newsgroups http://www.dejanews.com

Liszt Mailing Lists http://www.liszt.com

Companies http://www.companiesonline.com/

Edga http://www.sec.gov/edaux/searches.htm

Directories

Yahoo (directory http://www.yahoo.com

Yahoo Search Options http://search.yahoo.com/bin/search/options

Magellan (directory) http://www.mckinley.com

Magellan Search Options

Galaxy Professional Directory http://www.einet.net/

Galaxy Adv. Search http://www.einet.net/cgi-bin/wais-text-multi

Lycos A2Z Internet directory http://a2z.lycos.com/

Infoseek Directory http://www.infoseek.com/

Nerd World Subject Index http://www.nerdworld.com

Jump City (+ newsgroups) http://www.jumpcity.com/list-page.html

Your Personal Net http://www.ypn.com

Starting Point http://www.stpt.com/

Suite 101 http://www.suite101.com/

Brint: A Business Researchers Interest http://www.brint.com/interest.html

Martindale's Reference Center http://www-sci.lib.uci.edu/~martindale/Ref.html

The Mining Company Subject Site Guides http://miningco.com/

Top Site and Award Directories

Lycos Pointcom Top 5% http://www.pointcom.com/categories/

Netguide Live (go to Best of the Web) http://www.netguide.com

Librarian's Guide: Best Info on the Net http://www.sau.edu/CWIS/Internet/Wild/index.htm

Looksmart Directory http://www.looksmart.com

NBN News Editor Choice Awards http://nbnews.com/

Web Scout Best Link http://www.webscout.com

Cnet's Best of the Web http://www.cnet.com/Content/Reviews/Website/Pages/WS.categories.html

RoadKill Cafe's 175 Great Sites http://www.calweb.com/~roadkill/great.html

Digital Librarian Best of the Web http://www.servtech.com/public/mvail/home.html

TOP Web Site Lists

Web21 100 Hot Web Sites* http://www.web21.com/

The Web 100 http://www.web100.com/listings/all.html

WebCounter Top 100 http://www.digits.com/top/both_100.html

Zenation's Top 100 http://www.zenation.com/loto.htm

WebSide Story Top 1000 http://www.hitbox.com/wc/world2.html

Ziff-Davis' ZDNET http://www.zdnet.comCNEThttp:/www.cnet.com

ROBIN PEEK
Simmons College