New SLAC Web Information Manager
Appointed

Pat Kreitz, WWW-Coordinating Committee Chairperson
Dennis Wisinski, Manager, SCS Web Support Group

The very early days of the Web at SLAC have been history-making, as documented in Jean Deken's paper, "First in the Web, But Where Are the Pieces?" One of the central figures in the pioneering days of the Web at SLAC was, and has continued to be, Joan Winters.

Originally working solely as a volunteer, but later receiving the recognition of having the responsibility for SLAC's first home page officially included in her SCS job description, Joan has created and managed SLAC's top-level highlighted and detailed home pages and has written much of the support and instruction pages that exist in the SLAC Web-site. Many of the documents Joan has written and the ideas she has championed have resulted in a careful and consistent management of SLAC's top-level Web information. As part of her efforts, she helped ensure that both neophytes and sophisticated Web users could find the information they needed with ease. Joan's many years of dedicated work have laid a lasting foundation of solid Web management at SLAC.

One certainty in Web-land as well as the remainder of life at SLAC is that change is a constant. In order to assume new responsibilities within SCS, Joan Winters will no longer be managing SLAC's top-level Web. As of April 1, 1998, Ruth McDunn has assumed that responsibility. Ruth's time will be divided, with half of it remaining in Technical Publications as Information and Document Technology Specialist and half of it now reporting to the Chair of the Web Coordinating Committee as Web Information Manager. Ruth brings many strong technical and publication skills to her new role. She welcomes your comments and suggestions in the coming months as she settles into her new responsibilities.

Please continue to use the email address: www-admin@slac.stanford.edu to send all Web-related requests and any comments, problems, or requested changes to the top-level SLAC Web pages. Ruth and the SCS Web Group will be working together to ensure that Web requests and other Web-related concerns are quickly addressed by the appropriate person. If you wish to discuss a Web-related issue with Ruth in real-time, she can be reached at x2014.

For more information, please contact Pat Kreitz or Dennis Wisinski.

Last update: 1 April 1998
1 Introduction

Aim of this Document

This handbook, *Running A World Wide Web Service*, has been funded by the Advisory Group On Computer Graphics to provide support for UK academic institutions who wish to run a World Wide Web service. The objectives of the document are to provide the reader with:

An understanding of the history of the WWW.

Examples of how WWW is currently being used in a number of areas.

An understanding of the role of a WWW browser (client) and examples of some of the popular clients.

An understanding of the Hypertext Markup Language (HTML) and a review of popular HTML authoring tools, document converters and quality tools.

An understanding of the use of graphics on WWW and a review of some of the popular graphical tools.

Examples of some of the popular searching tools which are available on WWW and a review of some of the popular indexing tools.

An overview of WWW server software and a review of the main functions carried out by a server.

Examples of extensions to the basic WWW architecture, including the use of external browsers, running client-side and server-side applications.

An awareness of the legal and ethical aspects of providing a WWW service.

The handbook also gives a number of examples of the use of WWW in the following areas:

Running a campus (or community) wide information system (CWIS).

Teaching and learning.

Collaboration.

Libraries.

The handbook also provides pointers to a variety of sources of further information including:

WWW mailing lists.

National UK WWW services.

Target Audience

This document is intended primarily for the UK academic community. It should be suitable for computing service or administrative staff responsible for managing a World Wide Web service, and for academic staff who wish to run a departmental service.

2 About The World Wide Web

History

The World Wide Web (which is often referred to as W3, the Web or, as used in this document, WWW) is a distributed multimedia hypertext system. What is meant by this?

**Distributed:** information on WWW may be located on computer systems around the world.

**Multimedia:** the information held on WWW can include text, graphics, sound and even video.

**Hypertext:** access to the information is available using hypertext techniques, which typically involve using a mouse to select highlighted phrases or images. Once a phrase or image is selected it can result in information being retrieved from around the world.

The World Wide Web was initially developed by Tim Berners-Lee and Robert Cailliau of CERN Laboratories, Geneva to provide an infrastructure for particle physicists throughout Europe to share information. Since the physicists were located in various organisations and used a variety of computer systems and applications software
(Including various word processing and text markup programs for producing reports) the World Wide Web was developed using the client–server architecture, which ensured cross-platform portability.

Client–Server Architecture

The World Wide Web is based on the client–server architecture which is illustrated in Figure 1–1.

Click here for Picture.

Figure 1–1 WWW Client–Server Architecture.

The end user accesses the World Wide Web using a browser client, typically on a desktop machine such as a PC, Macintosh or Unix workstation. The client will display hypertext links in some manner, such as underlining the links. Selecting a link (by clicking a mouse button with a graphical client, typing the number following the link using a simple text–based client or using speech or foot pedals, for example, with browsers for disabled users) to send a request over the network (which could be a local network, a national network such as JANET, or over the global network which can be referred to as the Internet). The request is sent to a World Wide Web server, which is typically running on a powerful computer system. The server will retrieve the file which has been requested and deliver it to the client.

Once the client has started to retrieve the file it can display it on the local machine. If the client cannot display the file (many clients, for example, cannot view video clips) the client can pass the file on to an external viewer which can process the file.

This is a very simple overview of the WWW client–server architecture. Many other features are available: for example the server could send a message to the client, saying that the user is not authorised to access the file. However an understanding of this model will help you to see how the WWW can develop.

Early Browsers

One of the first browsers to be developed was the CERN command line browser. This can be accessed by using the command:

telnet telnet.w3.org

from a computer system which runs the telnet software. An example of use of the CERN command line browser is illustrated below.

telnet telnet.w3.org

Welcome to the World–Wide Web

THE WORLD–WIDE WEB

This is just one of many access points to the web, the universe of information available over networks. To follow references, just type the number then hit the return (enter) key.

The features you have by connecting to this telnet server are very primitive compared to the features you have when you run a W3 "client" program on your own computer. If you possibly can, please pick up a client for your platform to reduce the load on this service and experience the web in its full splendor.

For more information, select by number:

A list of available W3 client programs[1]
Everything about the W3 project[2]
Places to start exploring[3]
The First International WWW Conference[4]
This telnet service is provided by the WWW team at the European Particle Physics Laboratory known as CERN[5]
1–5, Up, Quit, or Help:

Figure 1–2 The CERN Command Line Browser.

Notice that in the CERN command line browser in order to select a hypertext link you need to type the number which follows the link.

The CERN command line browser is a very simple client. The first WWW browser was developed by Tim Berners-Lee, the father of the World Wide Web, for the NeXt system. However the NeXt hardware was not a commercial success and is no longer manufactured. One of the earliest graphical browsers was the Viola client which was developed for the X windows environment. Viola is illustrated in Figure 1–3.

Click here for Picture

Figure 1–3 The Viola Client.

Notice that in the Viola client the hypertext links are identified by the use of underlining.

The way in which hypertext links are displayed is a function of the client; they are not embedded in the underlying document. It is important to understand this in order to avoid writing documents which say "Type the number after the phrase", "Click on the underlined phrase" or "Click on the blue writing".

Growth In Popularity

As shown in Figure 1–4 use of the WWW has grown tremendously since 1993 (this file is available at the URL ftp://ftp.isoc.org/isoc/charts/networks.gif). Much of this growth in popularity was due to the release of browsers for the Windows, PC and Macintosh environments by the National Center for Supercomputing Applications (NCSA) at the University of Illinois.

Click here for Picture

Figure 1–4 Growth In Popularity of WWW.

Since CERN's remit was research in particle physics the WWW development team realized that they needed to involve other organizations in WWW development work. The involvement of NCSA in the WWW development programme resulted in the NCSA Mosaic for X, which was released in early 1993. An illustration of a pre-release version of Mosaic for X is shown in Figure 1–5.

Click here for Picture

Figure 1–5 A Pre-release Version Of NCSA Mosaic For X.

As can be seen from Figure 1–5 NCSA Mosaic for X provides access to a number of types of resources, including WAIS, Gopher, FTP, Usenet, Hytelnet, TeXinfo, X 500 and Whois resources. NCSA Mosaic was developed by a group of programmers at NCSA led by Marc Andreessen. NCSA Mosaic for X became such a success because:

It was a professional looking product from its initial release.

It used the Motif toolkit and was available on workstations platforms widely used in academia.

It made use of NCSA's expertise in scientific visualisation.

It introduced inline images.

In November 1993 NCSA released versions of Mosaic for the Microsoft Windows and Apple Macintosh environment. These browsers, which are freely available to the academic community, provided access to WWW to people who did not have access to Unix and X–Windows systems.

Terminology

The following terms have been used in this document:

Browser: An interactive program which is used to access information held on the World Wide Web.

Client: Often used as a synonym for browser. A client is the software which normally runs on the local desktop machine. The client sends requests to the server software.

Server: Software which is used to deliver information to a client. Can also refer to the computer system on which the server software is running.

The following terms will also be used in this document:

URL: Uniform Resource Locator. Can be regarded as the address of a file on the World Wide Web. It includes
the protocol (rules) for retrieving the file, the domain (name) of the computer system on which the server software runs and the file name to be retrieved. For example the URL http://info.cern.ch/TheProject.html uses the http protocol to retrieve the file TheProject.html from the computer called info.cern.ch


HTTP: Hypertext Transport Protocol. The protocol (set of rules) used to define the communications between the client and WWW server software.

Note that these terms are, for reasons of clarity, in some cases over–simplified.

Examples of Usage

A number of examples of how the World Wide Web is currently being used are given below.

Publishing Research Information

Figure 1–6 illustrates how CERN (the European Particle Physics Laboratory) makes its technical papers available on the World Wide Web. The URL for this technical paper is http://www1.cern.ch/ALICE/ENGINEERING/engineering.html

[Click here for Picture]

Figure 1–6 Scientific Information Held At CERN.

Campus Wide Information Systems

The Honolulu Community College Campus Wide Information System (CWIS) was the first multimedia CWIS on the World Wide Web. The URL for this CWIS is http://www.hcc.hawaii.edu/

[Click here for Picture]

Figure 1–7 The Honolulu Community College CWIS.

Teaching Applications


[Click here for Picture]

Figure 1–8 A Distributed Teaching Application.

Publicity

The School of Computer Studies at the University of Leeds was one of the first departments to use the multimedia capabilities of WWW to market its courses to potential students. The URL for this application is http://agora.leeds.ac.uk/WWW/MSc/MSc_text/leeds.html

[Click here for Picture]

Figure 1–9 Leeds University Postgraduate Prospectus.

Virtual Libraries

Many virtual libraries, art galleries and exhibitions are available on the World Wide Web. One of the first was the Vatican exhibition. The URL for this virtual exhibition is http://sunsite.unc.edu/expo/vatican/exhibit/Vatican.exhibit.html

[Click here for Picture]

Figure 1–10 The Vatican Exhibition.

Commercialisation Of WWW

The World Wide Web is increasingly being used by commercial companies. The URL for the Pizza Hut ordering service is http://www.pizzahut.com/

[Click here for Picture]

Figure 1–11 Commercial Applications On WWW.
3 World Wide Web Clients

In order to access the World Wide Web you will need to use a client (or browser). A wide range of clients are available for many different platforms; although the Mosaic client is very popular you should not think that Mosaic is the World Wide Web.

Publicly Available Clients

A number of clients are publicly available using the telnet protocol. These include:

telnet.w3.org

dir.mcc.ac.uk

rslox.ac.uk

lynx.mallbase.ac.uk

These clients can be accessed by giving the command telnet address for example telnet dir.mcc.ac.uk In some cases you will automatically be logged in, in other cases you must enter a username, which is often lynx.

An example of the use of the client at the Radcliffe Science Library at Oxford University is illustrated in Figure 3–1.

telnet rslox.ac.uk

Radcliffe Science Library & Bodleian Library WWW Server (p1 of 6)

RADCLIFFE SCIENCE LIBRARY & BODLEIAN LIBRARY WWW SERVER

UNIVERSITY OF OXFORD

[IMAGE]

Welcome! At present this WWW server is still feeling its way. This page is intended primarily as a starting point for Oxford users wishing to explore Internet services and information sources. From this home page you can also access some of our Local WWW applications which are for the most part still under development. For newcomers to the Web, one good introduction is Entering the World–Wide–Web: A Guide to Cyberspace by Kevin Hughes. Another is CERN’s WWW FAQ (list of Frequently Asked Questions).

Apologies to our regular Lynx users. We have phased out the old Lynx opening page and you will now commence with this one. If you would like to voice your opinions or your feelings, please feel free to use the comments form below.

-- press space for more, use arrow keys to move, 'p' for help, 'q' to quit

Figure 3–1 The Client At Radcliffe Science Library.

It should be noted that the organisations running these publicly available clients do not guarantee to provide the service on a long term basis.

Command Line Clients

The client illustrated in Figure 3–1 is a command–line client. Command–line clients typically run in a text–based operating system environment (e.g. DOS rather than Microsoft Windows or Unix rather than X Windows). Command line clients place less demands on the local computer system, but do not provide the ease–of–use or range of functionality provided by graphical clients.
Lynx

The most widely-used command line client is probably Lynx. Lynx was developed at the University of Kansas. Lynx was originally developed for Unix. An example of the Unix implementation is illustrated in Figure 3-1. Lynx has been ported to the MS DOS environment. DosLynx, as the implementation is known, will run on a PC with 512 K of RAM, running MS DOS 3 or above. It provides access to the World Wide Web from an entry level PC which has the appropriate networking capability. DosLynx is illustrated in Figure 3-2.

Click here for Picture

Figure 3-2 DOS Lynx

Availability

The Lynx browser software is available at the URL ftp://ftp2.cc.ukans.edu/pub/Web/. In the UK it is also available at the URL ftp://src.doc.ic.ac.uk/packages/WWW/lynx

The DosLynx software is available at the URL ftp://ftp2.cc.ukans.edu/pub/WWW/DosLynx/

Details of the system requirements for DosLynx are available at the URL ftp://ftp2.cc.ukans.edu/pub/WWW/DosLynx/readme.htm. A Listserv mailing list exists at the address Dos lynx-Dev@ukanaix.cc.ukans.edu for the distribution of DosLynx related information, updates, and development discussion. To subscribe send a request to listserv@ukanaix.cc.ukans.edu to be added to the list. All new releases will be announced on this list.

NCSA Clients

The NCSA Mosaic client is available for the X Windows, Microsoft Windows and Apple Macintosh environments.

NCSA Mosaic For X

Although it was not the first graphics browser, NCSA Mosaic For X helped to popularise the Web. At the time of writing version 2.4 is available, although a beta version of 2.5 is also available, which includes support for hierarchical hotlists.

Click here for Picture

Figure 3-3 NCSA Mosaic For X.

NCSA Mosaic For Windows and the Macintosh

If NCSA Mosaic For X helped to popularise the Web, NCSA Mosaic For Windows and for the Macintosh made it available to large numbers of people.

Click here for Picture Click here for Picture

Figure 3-4 NCSA Mosaic For Windows and the Macintosh.

Availability

The NCSA Mosaic browser software for the X, Microsoft Windows and Apple Macintosh platforms is available at the URL ftp://ftp.ncsa.uiuc.edu/pub/Web/. In the UK it is also available at the URL ftp://src.doc.ic.ac.uk/packages/WWW/Mosaic/

Cello

Cello was one of the first WWW browsers to be developed for the Microsoft Windows environment. It was written by Thomas R Bruce of the Legal Information Institute, Cornell University.

Click here for Picture

Figure 3-5 The Cello Browser.

Availability

The Cello browser software for Microsoft Windows is available at the URL ftp://ftp.law.cornell.edu/pub/II/Cello/

HNet Clients

HNet have developed the WinWeb and MacWeb clients, which are based on the NCSA Mosaic source code.
Figure 3–6 The WinWeb and MacWeb Browsers.

Availability
The EINet browsers software for the X, Microsoft Windows and Apple Macintosh environments are available at the URL ftp://ftp.einnet.net/einnet/

Netscape Clients

Mosaic Communications Corporation (MCOM) was set up by Jim Clark, founder of Silicon Graphics. MCOM recruited the developers of NCSA Mosaic to develop a WWW browser. A beta release of Netscape was released in October 1994. It generated a tremendous amount of interest, because of its speed and functionality. However it also caused concern, since it included extensions to the HTML standard which had not been part of the HTML standardisation process.

Note that for legal reasons MCOM was renamed Netscape in November 1994.

Figure 3–7 The Netscape Browser for Windows and the Macintosh.

Availability
The Netscape browser software for the X, Microsoft Windows and Apple Macintosh environments is available at the URL ftp://ftp.netscape.com/ In the UK it is also available at the URL ftp://src.doc.ic.ac.uk/packages/WWW/Netscape/ Further information is available from the URL http://home.netscape.com/home/welcome.html

AirMosaic

Air Mosaic is another commercial browser which is based on the NCSA Mosaic source code.

Figure 3–8 The AirMosaic Browser.

Availability
An evaluation copy of the AirMosaic browser software for the X, Microsoft Windows and Apple Macintosh environments is available at the URL ftp://ftp.spry.com/demo/ Further information is available at the URL http://www.spry.com/

GWHIS

GWHIS is a commercial WWW browser, marketed by Quadralay. GWHIS (Global-Wide Help and Information System) consists of a WWW browser, an application program interface (API) for integrating GWHIS into applications and a search engine.

Figure 3–9 The GWHIS Browser For X Windows.

Availability
An evaluation copy of the GWHIS browser software for the X, Microsoft Windows and Apple Macintosh environments is available at the URL ftp://ftp.quadralay.com/pub/gwhis Further information is available at the URL http://www.quadralay.com/

Other Browsers

Many other browsers are available or are currently being developed. Some of the browsers are aimed at the business community. Of particular interest to the academic community are the Internet browsers which are being developed by IBM, Microsoft (for inclusion with Windows 95) and Novell.

Figure 3–10? The Warp OS/2 Browser.

Future Developments

A browser known as Arena is currently being developed which will handle HTML 3. HTML 3 is a new version of HTML which contains a number of facilities which are not available in HTML 2 including table handling and mathematical formulae.
Click here for Picture

Figure 3–9 The Arena Browser.

Availability

Arena is currently a beta program. It can be obtained from the URL ftp://ftp.w3.org/ In the UK it is available at the URL ftp://src.doc.ic.ac.uk/packages/cern/arena/

Conclusions

Which is the best browser? There is no longer a simple answer to this. The growth in the number of browsers, the different licensing arrangements and different areas they address is making it difficult to adopt an institutional policy on choosing a browser. At the time of writing the Netscape browser looks very attractive. However it was developed primarily to address the needs of commercial users, many of whom requested greater control over the appearance of HTML pages, in order to reflect a corporate identity. Will Netscape, however, be as quick to support mathematical equations, which will be of interest to most academic institutes? Will it be the best browser for providing control over external applications – an area which is likely to be of interest to academics who wish to develop distributed teaching materials?

Perhaps the only conclusion to be made at this point is that academic institutions should avoid being locked in to a particular browser.

4 HTML

About HTML

Native documents on the World Wide Web or written in HTML, the Hypertext Markup Language, HTML, define the structural elements in a document (such as headers, citations, addresses, etc.), layout information (bold and italics), the use of inline graphics together with the ability to provide hypertext links.

A simple HTML document is illustrated in Figure 4–1.

<TITLE>The World Wide Web</TITLE>

<H1>About The World Wide Web</H1>

P>The World Wide Web is a EM>distributed multimedia hypertext</EM> system.</P>

Figure 4–1 A Simple HTML Document.

Structural elements in the document are identified by start and end tags. For example the <TITLE> and </TITLE> tag is used to specify the title of the document, which is often displayed by a client. The <H1> and </H1> tag is used to define the first level heading. Clients will normally display headers differently from the body text: for example, a graphical client could display the header using a larger or different font, whereas a text–based client could display a header as centred text or in all capitals.

Figure 4–1 also illustrates the <EM> container. Text held in the container (which is defined by the <EM> start tag and the </EM> end tag) will be emphasised in some way. A graphical browser could render the emphasis by displayed the text in italics, whereas a browser with audio capabilities for the visually impaired could render the emphasis by a change in the tone of the voice output.

Figure 4–1 also shows the paragraph container. It is important to understand that the <P> tag is part of a paragraph container and is not a paragraph separator (as many people mistakenly believe). If the <P> is not used the existence of the next <P> tag will imply a </P>. In future versions of HTML it will be possible to specify paragraph attributes, such as <P ALIGN=Centred>.

Although browsers will display the HTML document shown in Figure 4–1 for reasons of performance and upwards compatibility it is strongly recommended that HTML documents contain additional elements including the <HTML>, <HEAD> and <BODY> tags, as shown in Figure 4–2.

<HTML>

<HEAD>

<TITLE>The World Wide Web</TITLE>

</HEAD>

<BODY>

<H1>About The World Wide Web</H1>
Information about the World Wide Web is available at CERN.

Figure 4-2 A Simple HTML Document.

The <HTML> container is used to define the extent of the HTML document. Within the HTML document there are two other containers: <HEAD> and <BODY>. The <HEAD> container provides information about the document itself. This can include the title of the document, as illustrated, copyright information, keywords and expiry dates (for use by caching software). It is important to make use of the <HEAD> tag since, for example, an automatic indexing program which wishes to index the title of HTML documents can use the HTTP protocol to retrieve only the information contained in the HEAD container. If the HEAD container is not present the entire document will be retrieved, which will place unnecessary extra load on the server and on the network.

Figure 4-2 also illustrates the use of the anchor <A> container. This tag is used to provide hypertext links. In the example the text at CERN which is contained between the <A> and </A> tags will be highlighted in some way by the browser. Selecting this highlighted phrases will cause the client to send a request for http://info.cern.ch/hypertext/WWW/TheProject.html This request will use the http protocol and will be sent to the server running on the system at info.cern.ch

HTML Authoring Tools

Initially information providers on the World Wide Web used standard editors such as vi and emacs to create HTML documents. As WWW matured authoring tools were developed to assist information providers.

HTML Assistant

HTML Assistant is a simple authoring tool which can be used to create and edit HTML documents. Frequently Asked Questions about HTML Assistant is available at the URL http://cs.dal.ca/ftp/htmilassht/htmlfaq.html HTML Assistant is available at the URL ftp://ftp.cs.cornell.edu/pub/pc/win3/misc In the UK it is available at the URL ftp://src.doc.ic.ac.uk/packages/WWW/tools/editing/ms-windows/html-assistant

Click here for Picture.

Figure 4-3 HTML Assistant.

HTML Hyperedit

HTML Hyperedit is a Toolbook application which is available for the Microsoft Windows environment. As well as providing an environment for producing HTML documents, it also a tutorial which gives an introduction to HTML. HTML Hyperedit is available at the URL ftp://info.curtin.edu.au/pub/internet/mswindows/hyperedit In the UK it is available at the URL ftp://src.doc.ic.ac.uk/packages/WWW/tools/editing/ms-windows/win--htmledit

Click here for Picture.

Figure 4-4 HTML HyperEdit

HTML Writer

HTML Writer is a simple authoring tool which can be used to create HTML documents.

Click here for Picture.

Figure 4-5 HTML HyperWriter.

HTMLEd

HTMLEd is another simple authoring tool which can be used to create HTML documents. In the UK it is available at the URL ftp://src.doc.ic.ac.uk/packages/WWW/tools/editing/ms-windows/

Click here for Picture.

Figure 4-6 HTMLEd.

Word Processing Tools

HTML Assistant and HTML Hyperedit and self-contained authoring tools. A number of authoring tools have been developed which work within a word processing environment. These tools are normally implemented as macros for popular word processing packages, such as Word For Windows or WordPerfect.
Word processing tools have the advantage that they provide a consistent environment for existing users of word processors. However, they do have their disadvantages. Because they are normally implemented as macros, they can be very slow, especially when used with large or complicated documents. There is also a danger that HTML markup which is embedded as hidden text could cause conflicts with other word processing tools if, for example, the word processed document was used by other users.

**GT_HTML**

One of the first word processing macros which could be used to create HTML documents was the GT_HTML macro. This macro, written for Word For Windows, was developed at the Georgia Technical Research Institute. In the UK the software is available at the URL ftp://src.doc.ic.ac.uk/packages/WWW/tools/editing/macros/ms-winword

[Click here for Picture]

**Figure 4–7 The GT_HTML Macro.**

**CU_HTML**

CU_HTML is a template designed to work with Word For Windows. The template was written by Anton Lam (mailto:anton-lam@cuhk.hk) The software is available at the URL ftp://ftp.cuhk.hk/pub/ww/windows/cuhk

[Click here for Picture]

**Figure 4–8 The CU_HTML Macro.**

**ANT_HTML**

ANT_HTML is a template designed to work within Word For Windows 6.0. The template was written by Jill Swift (mailto:jswift@freenet.fsu.edu) The software is available at the URL ftp://ftp.einet.net/einet/pc/ANT_HTML.ZIP

[Click here for Picture]

**Figure 4–9 The ANT_HTML Macro.**

**Browser Editing Tools**

Another approach to editing HTML documents is provided by browsers which are integrated with an editing tools. The Arena browser enables an external editor to be invoked to edit the displayed HTML document. Figure 4–10 illustrates the Arena browser used in conjunction with the Emacs editor.

[Click here for Picture]

**Figure 4–10 Editing A Document From Arena.**

**HTML Document Conversion Tools**

Authoring tools are normally used to create new HTML documents. Document conversion tools can be used to convert existing documents to HTML format.

**LaTeXtohtml**

One of the first sophisticated document conversion tools to be developed was the LaTeXtohtml conversion program. This program was written by Nikos Drakos, Computer Based Learning Unit, University of Leeds. It provides a wide range of features, and sets the standard for document converters. The facilities it provides include:

- Automatic generation of hyperlinks at user-defined section breaks.
- Automatic conversion of graphics and formulae to embedded graphics.

Figure 4–11 illustrates a document which has been converted by the LaTeXtohtml conversion program.

[Click here for Picture]

**Figure 4–11 A Document Converted Using LaTeXtohtml.**

It is often possible to identify documents which have been converted using LaTeXtohtml is available at the URL ftp://src.doc.ic.ac.uk/packages/WWW/tools/converters/latex2html

**RTFtohtml**

The RTFtohtml conversion program enables RTF files (which can be produced by word processing packages such as Word For Windows) to be converted to HTML. The program was written by Chris Hector, Cray, based on RTF
parsing software developed by Paul DuBois.

RTFtohtml is available as a command line tool for a number of Unix platforms. In addition an Apple Macintosh implementation is available. A beta version of an MSDOS implementation was announced in November 1994.

An extension of the RTFtohtml program is known as RTFtohtml. This provides a number of additional features, including creation of hypertext links at user defined section breaks. Figure 4–12 illustrates a document on Exploring The World Wide Web Using Mosaic For Windows which is available at the URL http://www.leeds.ac.uk/acs/docs/tut50/tut50.html

>>

Exploring The World Wide Web Using Mosaic For Windows (TUT 50)

Task 1 Launching Mosaic For Windows

Task 2 Using Mosaic For Windows

Task 2 Using the World Wide Web

...
forms and a CGI script which runs a HTML validation program. The service can be used to check HTML syntax by entering the HTML markup to be checked. It can also be used to check an existing HTML document by entering the URL of the document.

Click here for Picture.

Figure 4–14 HTML Validation Service.

A variation on this service is available at the URL: http://www.cc.gatech.edu/grads//Kipp.Jones/Validation/validation-form.html

These services make of the sgmls validation program.

The software can be installed on your local Unix system. It is available at the URL, ftp://ftp.hal.com/pub/CGI/check-html.tar.Z

HTML Check Toolkit

The HTML Check Toolkit is another HTML validation program. The software can be installed using a WWW browser. The installation service, illustrated below, is based on the EIT Webmaster Starter’s Kit. HTML Check Toolkit is available at the URL, http://www.hal.com/~markg/halSoft/html–check/

Click here for Picture.

Figure 4–14 Installing The Check_HTML Script.

SGMLS

sgmls is a tool which can be used to validate SGML documents. It is available at the URL, ftp://sgml1.ex.ac.uk/pub/SGML/sgmls/ sgmls is used in a number of HTML validation services, such as those mentioned above.

Forms

Creating A Form

A form consists of areas of the screen in which the user can input data. The data is sent to the HTTP server, which can run a script or program to process the data in some way. One common use of forms is to provide feedback on a WWW service. Input to the form can be emailed to the service administrator. Forms can also be used to input search criteria to be input to a search engine, or to specify parameters for distributed teaching and learning services.

A form is defined by the <FORM > tag. The <FORM > tag has the syntax

<FORM METHOD="method" ACTION="url"/>

For example:

<FORM METHOD="post" ACTION="http://leeds.ac.uk/ucs/cgi-bin/myscript">

will send the input data to be processed by myscript.

An example of a form is shown below:

<TITLE>Fill–Out Form Example #7</TITLE>

<H1>Fill–Out Form Example #7</H1>

This is another fill–out form example, with toggle buttons. <P>

<HR>

<FORM METHOD="POST" ACTION="http://boohoo.ncsa.uiuc.edu/htbin–post/post–query">

<H2>Godzilla’s Pizza — Internet Delivery Service, Part II</H2>

Type in your street address: <INPUT NAME="address"/><P>

Type in your phone number: <INPUT NAME="phone"/><P>

Which toppings would you like? <OL>
<LI> <INPUT TYPE="checkbox" NAME="topping" VALUE="pepperoni"> Pepperoni.

<LI> <INPUT TYPE="checkbox" NAME="topping" VALUE="sausage"> Sausage.

<LI> <INPUT TYPE="checkbox" NAME="topping" VALUE="anchovies"> Anchovies.

<OL>
How would you like to pay? Choose any one of the following: <P>

<OL>
<LI> <INPUT TYPE="radio" NAME="paymethod" VALUE="cash" CHECKED> Cash.

<LI> <INPUT TYPE="radio" NAME="paymethod" VALUE="check"> Check.

<LI> Credit card:<A>

<UL>
<LI> <INPUT TYPE="radio" NAME="paymethod" VALUE="mastercard"> Mastercard.

<LI> <INPUT TYPE="radio" NAME="paymethod" VALUE="visa"> Visa.

<LI> <INPUT TYPE="radio" NAME="paymethod" VALUE="americanexpress"> American Express.

</UL>

<OL>
Would you like the driver to call before leaving the store? <P>

<DL>
<DD> <INPUT TYPE="radio" NAME="callfirst" VALUE="yes" CHECKED> <I>Yes.</I>

<DD> <INPUT TYPE="radio" NAME="callfirst" VALUE="no"> <I>No.</I>

</DL>

To order your pizza, press this button: <INPUT TYPE="submit" VALUE="Order Pizza">. <P>

</FORM></FIGURE>

**Figure 4-15 HTML Document Defining A Form.**

This example is available at the URL
http://www.ncsa.uiuc.edu/SDG/Software/Mosaic/Docs/fill-out-forms/example-7.html

The way in which form is displayed is illustrated below.

Click here for Picture

**Figure 4-16 Using A Form.**

**Processing A Form**

Once the form is submitted the data which has been entered is appended to the end of the URL given in the ACTION attribute of the FORM tag. This information is then processed by the script.

**Further Information About Forms**

Forms tutorials are available at the URL:
http://www.ncsa.uiuc.edu/SDG/Software/Mosaic/Docs/fill-out-forms/overview.html,
http://boodoo.ncsa.uiuc.edu/docs/cgi/forms.html and http://kuhttp.cc.ukans.edu/info/forms/forms-intro.html
Writing Style

Writing styles for WWW documents are still developing. However there are a number of guidelines which can be provided:

Avoid saying "Click on the option". This is browser specific: with a command-line browser, such as Lynx, it is not possible to click to select hypertext links.

Avoid saying "Select here".

Avoid saying "Back to University Home Page". It the reader has gone directly to the page, the word back is not appropriate.

Avoid saying "University Home Page". Which University? Try to give the reader some context to where they are in cyberspace.

Finding Out More About HTML

This document not provide an in–depth tutorial on HTML. Many WWW resources are available which give details on writing HTML. Some of these are listed below:

A Beginner’s Guide To HTML is available at the URL http://www.ics.uci.edu/~soboroff/begin.html

The Elements Of HTML Style is available at the URL http://bookweb.cwis.uci.edu:8042/Staff/StyleGuide.html

HTML Documentation is available at the URL http://www.utm.utoronto.ca/HTML/docs/NewHTML/intro.html


Marc Andreessen's HTML Primer is available at the URL http://www.netscape.com/communicator/primers.html

Peter Flynn's Online Guide To HTML which is available at the URL http://www.ucc.ie/info/net/html/index.html

How to Write HTML files is available at the URL http://curia.ucc.ie/info/net/html/doc.html

Spinning the Web: An Introduction to HTML is available at the URL http://scholar.lib.vt.edu/reports/soasia–slides/HTML–Intro.html

Composing Good HTML is available at the URL http://www.williamette.edu/html–composition/strict–html.html

A Crash Course in HTML is available at the URL http://www.ziff.com/~camon/crash_course.html

In addition to these documents the following resources are also available:

Gabriel White (mailto:gabriel@werple.apana.org.au) has written a review of a number of Microsoft Windows HTML authoring tools which is available at the URL http://werple.apana.org.au/~gabriel/html–editors/index.html

Dan Connolly’s HTML Design Notebook is available at the URL http://www.html.com/%7Econnolly/html~design.html

A list of HTML tools is available at the URL http://info.cern.ch/hypertext/WWW/Tools/Filters.html

The HTML specification is available at the URL http://www.html.com/%7Econnolly/html~spec.html

5 Graphics

The World Wide Web is, of course, a graphical system. This section describes how graphical objects can be incorporated in an HTML document, how external graphical files can be used and how to create and use interactive maps. The section also considers the performance aspects of using graphics.

HTML Graphical Tags

Inline images are defined in an HTML document using the <IMG> tag. For example:

<IMG SRC="portrait.gif">

The full syntax for the <IMG> tag is:

<IMG SRC="source file" ALT="textual description" ALIGN="option">

The SRC attribute is used to specify the URL of the graphical file. At the time of writing graphical files should
normally be in GIF format, although the XBM format is also supported. The SRC attribute is mandatory.

The ALT attribute is used to specify text which should be described by a browser which cannot display graphics, or a browser which has the display of inline images switched off. Use of the ALT attribute is highly recommended.

The ALIGN attribute can take the values TOP, MIDDLE or BOTTOM. It is used to define whether the top, middle or bottom of the graphic should be aligned with the text. Use of the ALIGN attribute is optional.

Using External Browsers

You can use the <A> anchor tag to refer to a graphical file. When the link is selected the graphical file is normally passed to a graphical viewer (such as xv or LVIEW) for displaying.

One common use of the <A> tag is to provide a link to a large colour graphic from a small thumbnail image. For example:

<A HREF="full-image.jpeg"><IMG SRC="thumbnail.gif" ALT="Portrait of John Smith"></A>

It is also possible to use this technique to provide links from thumbnail images to video clips. For example:

<A HREF="fluidflow.mpeg"><IMG SRC="fluidflow.thumb.gif" ALT="Video clip of fluid flow"></A>

Graphical Tools

Mapedit

Mapedit is an editor for creating imagemap files. Image map files are a feature of NCSA and CERN servers; they enable you to turn a GIF image into a clickable map by designating areas using polygons and circles within the GIF and specifying a destination URL for each area. The software is not public domain. Commercial users must pay a licence fee, non-profit and educational users are asked to send the author a postcard. The software is available from the URL ftp://sunsite.unc.edu/pub/packages/info/graphics/WWW/tools/mapedit. In the UK it is available at the URL ftp://src.doc.ic.ac.uk/packages/WWW/mapedit. Mapedit was written by Thomas Boutell (mailto:btouell@netcom.com).

Click here for Picture.

Figure 5-1 MapEdit.

Paintshop Pro

Paintshop Pro is an example of a tool which can be used to manipulate graphics files for use on WWW. Paintshop Pro can be used to convert file formats, to reduce colour depth, to convert colours.

Click here for Picture.

Figure 5-2 Paintshop Pro.

The image being manipulated by Paintshop Pro contains information for 256 colour (as shown in the bottom left of the screen). The colour depth of the image should be reduced to decrease the size of the file, and thus the network traffic when the image is retrieved on WWW.

Other Graphical Tools

San Diego Supercomputer Center’s intools package converts many file formats, including GIFs.

ImageMagick is a multi-purpose raster converter and manipulation package. The convert program handles many file formats including GIF. The software is available at the URL ftp://ftp.x.org/contrib

Graphics Workshop from Alchemy Mindworks is a DOS package for converting graphical files.

Appropriate Use Of Graphics

Novice information providers may be tempted to fill their HTML documents with inline graphical images. More experienced computer users will remember the large numbers of poorly designed paper documents which were produced once desktop publishing packages became widely used.

Before making use of graphics you should consider the following points:

Will the image look attractive and be readable on computers with monochrome or limited colour capabilities?

Since many computers currently used have limited colour maps, images which contain many colours may not be displayed correctly.
Since the http protocol needs to reestablish a connection for every image, using many small images (such as coloured balls for bullets) can be inefficient.

The use of images to compensate for deficiencies in the current version of HTML can limit the effectiveness of browsers for the disabled (e.g. a browser for the visually impaired which provides audio output will not be able to handle graphical instructions and navigational aids).

The use of the same image (such as a logo) but held on different departmental servers results in unnecessary traffic, since the image will not be cached in the client.

Excessive use of graphics can act as a barrier to people in developing countries who may have to pay for their network traffic.

Look at the following URLs. See how long it takes for the information to be delivered. Note that if you retry the URL it is likely to be quicker if it is cached by your client or by your server (if your server supports caching).

http://www.lib.virginia.edu/

http://www.leeds.ac.uk/ucs/people/BKelly/uniras94/uk_logos.html

Further Information

A tutorial on imagemaps is available at the URL
http://winternute.ncsa.uiuc.edu:8080/map-tutorial/image-maps.html

A good example of use of graphics is the Xerox Parc Map viewer which is available at the URL
http://pubweb.parc.xerox.com/map

6 Searching And Indexing

The tremendous growth in the numbers and extent of information services on WWW has made net-surfing an ineffective way of finding useful information. Fortunately sophisticated indexing tools are being developed. Figure 6-1 shows a page which contains pointers to a number of searching tools.

Click here for Picture

Figure 6-1 A Collection Of WWW Search Engines.

A collection of WWW search engines is available at the URL http://cat.unige.ch/meta-index.html

Robots, Spiders and Worms

During 1993 many WWW users discovered resources by net-surfing: going to one WWW server, exploring what was available, and then following links to other WWW servers. A number of software developers produced software which automated this process, so that a program went from server to server, indexing information, such as contents of the <TITLE> tag or the contents of server home pages. Such programs became known as robots or spiders; one robot was called WWW, the World Wide Web Worm.

There are a number of problems with this approach to global indexing:

Server performance When a robot arrives at your server it can place a heavy load on the server.

Network performance Robots can place heavy loads on the network infrastructure.

Quality Robots index all files. They are unable to differentiate between valuable information resources (such as a collection of research reports) and low quality, transient information (such as an undergraduate’s personal interests).

Maintenance The robot will not know if an information source that has been indexed is withdrawn from service.

A number of these issues have been addressed. Martijn Koster’s Guidelines For Robots, which is available at the URL http://web.nexor.co.uk/mak/doc/robots/robots.html provides guidelines for developers of robots.

A list of robots is kept at the URL http://web.nexor.co.uk/mak/doc/robots/active.html

Aliweb

Aliweb (Archie Like Indexing In The Web) provides another approach to the indexing of WWW resources. With Aliweb each site is responsible for indexing files. The server administrator is responsible for choosing the files to be indexed.

Further information about Aliweb is available at the URL http://web.nexor.co.uk/aliweb/doc/aliweb.html. The paper ALIWEB – Archie-Like Indexing In The Web, which was presented at the WWW 94 conference in CERN.
SWISH

SWISH, which stands for Simple Web Indexing System for Humans, was announced on 16 November 1994. It is a program that allows you to index your Web site and search for files using keywords in a fast and easy manner. Documentation is available at the URL http://www.eit.com/software/swish/swish.html. The software is available at the URL ftp://ftp.eit.com/pub/web/software/swish/

WAIS

WAIS (Wide Area Information Server) is another mechanism for indexing resources. WAIS is used by the Computing Service, University of Leeds to index its documents and newsletters. An example of how the WAIS server and WAIS indexing software is used is given below.

The command:

waissv -p 210 -d /apps/info/WWW/WAIS

is used to start the WAIS server software. The -p 210 argument specifies the name of the port on which the server runs while the -d argument gives the name of the directory which will contain WAIS databases. Note that since the WAIS server will normally be running continuously it will normally be initiated by the system administrator.

Newsletters are indexed by giving the command

waissv -p 210 -d /apps/info/WWW/ucs/newsletter/wais-sources/computing--service--newsletter

HTML *html

The name of the WAIS database is computing--service--newsletter. This long name is used since a single directory is used for all WAIS databases — it will save confusion if other departments wish to index their own departmental newsletters.

The WAIS database can be accessed by a dedicated WAIS client or by a WWW browser which contains support for the WAIS protocol. The WAIS database can be accessed by giving the URL

waissv://www.leeds.ac.uk/computing--service--newsletter

WAIS Utilities

A number of utilities are available which can post-process the output from WAIS.

wais.pl is a CGI script which is distributed with the NCSA httpd server.

Son of wais.pl is a CGI script which is based on the wais.pl script.

SFGate is a CGI script which interfaces to WAIS servers. SFGate provides a forms interface which can be used to access a number of WAIS databases. It is available at the URL
http://hs6-1--www.informatik.uni--dortmund.de/SFGate/SFGate.html A demonstration is available at the URL
http://hs6-1--www.informatik.uni--dortmund.de/SFGate/multiple.html

wwwais is a small ANSI C program that acts as gateway between waisq or waissearch (programs that search WAIS indexes) and a forms-capable World-Wide Web browser. With the freely distributable freeWAIS package, this program, and your local Web site, you can:

Create searchable databases of the information on your Web site

Allow users to search multiple databases via their Web browser with customizable options

Create a custom pop-up menu of servers to search through

Produce hypertext search results, with file information and links directly to the relevant HTML documents

Retrieve WAIS source descriptions and files

Specify URLs and filters to map results to

Only allow users from certain sites to search documents

Documentation is at the URL http://www.eit.com/software/wwwais/wwwais.html

You can FTP the source and related files from the URL ftp://ftp.eit.com/pub/web/software/wwwais/

You can see how it looks at the URL http://www.eit.com/cgi--bin/wwwais

A WAIS Application
One interesting application of the use of WAIS is the multimedia archive prototype developed by Andy Walker, formerly of the CBL/Multimedia Unit, University of Leeds. The prototype was developed to investigate the feasibility of providing a archive of multimedia objects for use in CBI applications by members of the University of Leeds.

A directory is created for each multimedia object. The directory contains the multimedia object itself (e.g. a graphical file, video clip or sound file) together with a keyword file which describes the object. The keyword files are indexed using WAIS. A WWW browser which supports forms is used to run a CGI script. The CGI script invokes the waisq command to search the WAIS database. The output from waisq is then used to create an HTML file which contains a pointer to a thumbnail image of matching multimedia objects.

Click here for Picture

Figure 6-2 Multimedia Archive

Which WAIS?

A number of WAIS servers are available. The freeWAIS software is currently used at the University of Leeds. This software is maintained by CNIDR, the Clearinghouse For Networked Information Discovery and Retrieval. The freeWAIS software, however, is based on the 1988 version of the Z39.50 protocol. An implementation of WAIS based on the 1992 version of Z39.50 is also believed to be available from CNIDR.

freeWAIS-sf is an implementation of WAIS developed at Dortmund University, available at the URL
ftp://isb-informatik.uni-dortmund.de/pub/wais/freeWAIS-0.2-sf-beta.tar.gz

The author, Ulrich Pfeifer, sent the following comment to the comp.infosystems.wais Usenet news group on 12 July 1994:

I have invested by far more time I intended to, to make freeWAIS-sf

stable and portable. To quote Larry Wall: "I have taken too much time from my employer, my family, my career and my health ... ."

Now I have to focus my efforts on getting my PhD. Tung has to finish his Diploma thesis.

At the time of writing continuing support and development for freeWAIS-sf is uncertain.

Further Information

A tutorial on Mosaic and WAIS is available at the URL
http://wintumble.ncsa.uiuc.edu/8080/wais-tutorial/wais.html

A WAIS overview is available at the URL http://info.cern.ch/hypertext/Products/wais/sources/Overview.html

See also http://wais.com/

A list of resources about the Z39.50 information discovery protocol is available at the URL
http://ds.internic.net/z3950/z3950.html

7 WWW Servers

If you wish to make information available you will need to run a WWW server. The server software is known as httpd – the hypertext transport protocol daemon. Just as their many WWW browsers available there are also many servers, including ones for Unix, MS Windows, Windows NT and the Apple Macintosh.

This section gives an example of how to install and run a server for the Microsoft Windows environment. The section then goes on to illustrate a number of server management issues which are based on the CERN server for the Unix platform.

Example of Installing A Server

An example illustrating how easy it is to install a WWW server is given below. The example assumes that you have access to a networked PC.

Connect to the NCSA server software from the anonymous FTP server at ftp.ncsa.uiuc.edu. Then change directory to /Web/httpd/Uni/ncsa_httpd/contrib/winhtml. Finally retrieve the file whpp13p1.zip An example of how to do this using the FTP software is illustrated below.

ftp src.doc.ic.ac.uk

image
cd /Web/ncsa/httpd/Windows/winhttpd

get whttp13p1.zip

Create a directory called C:HTTPD on the C: drive of your PC. Then uncompress the file by giving the command:

PKZUNZIP -D WHHTTP13P1.ZIP

The -D option will preserve the directory structure from the compressed file.

Run Microsoft Windows and create a program icon using the New option on the File menu. The icon should point to the file C:\HTTPD\HTTPD.EXE.

Set the time zone in the configuration file so that T2=GMT.

Run the server program. The window shown below should be displayed.

Click here for Picture.

Figure 7–1 Running The Windows HTTPD Server.

Run a World Wide Web browser and then enter a URL containing the IP address of your PC. For example if your PC has an IP address of 192.11.1.1 you should enter the address:

http://192.11.1.1/

The following diagram illustrates how NCSA Mosaic for X can be used to access a server running on a PC

Click here for Picture.

Figure 7–2 Accessing Your HTTPD Server.

This example is meant to illustrate the installation of a WWW server. In practise the server software is likely to run on a more robust system than a PC running MS DOS, such as a Unix or Windows NT system.

Server Configuration Files

World Wide Web server software will normally have a configuration file which is used to:

Specify an area of the file store on the server machine from which files can be retrieved.

Restrict access to certain files.

Specify caching, which enables remote files to be stored on the local server for a period of time.

Act as a proxy gateway, which provides a means for unauthorised machines to access resources on WWW.

As WWW develops, additional features will be provided in the server software and the configuration files are likely to grow in complexity. An example of a simple configuration file is shown below.

map / file/apps/WWW/homepage.html

map /* file/apps/WWW/*

pass file/apps/WWW/*

fail *

Figure 7–3 A Simple httpd.conf Configuration File

Figure 7–3 shows a simple configuration file for the CERN httpd server. Line 2 specifies that files located under the directory /apps/WWW should be available to the WWW server software. Line 1 specifies that file /apps/WWW/homepage.html is the default file to be displayed when the WWW server is accessed.

protection prot–proxy { # Part 1

serverid www.leeds.ac.uk

mask @(129.11. * *)

}

protect http:* prot–proxy # Part 2
protect gopher:* prot-proxy
protect ftp:* prot-proxy
protect wais:* prot-proxy
pass http:* # Part 3
pass gopher:*
pass ftp:*
pass wais:*

# Part 4
Exec /cgi-bin/ucs/* /apps/WWW/cgi-bin/ucs/*
Exec /cgi-bin/bionet/* /apps/WWW/cgi-bin/bionet/*
Exec /cgi-bin/bmb/* /apps/WWW/cgi-bin/bmb/*

# Part 5
map / file:/apps/WWW/homepage.html
map /* file:/apps/WWW/*
pass file:/apps/WWW/*
fail *

AccessLog /var/adm/httpd.log # Part 6
LogFormat Common
LogTime LocalTime
Caching On # Part 7
CacheRoot /usr/info/WWW_cache
CacheSize 300
CacheAccessLog /var/adm/httpd_cache.log

# Part 8
# Delete files from cache after specified number of days
CacheClean http:* 10 Days
CacheClean gopher:* 10 Days
CacheClean wais:* 10 Days
CacheClean ftp:* 10 Days
# Don't cache local files # Part 8
NoCaching http://*.leeds.ac.uk/*
# If a file hasn't been accessed within the last specified
# number of days delete from cache
CacheUnused * 5 days
CacheUnused http://info.cern.ch/* 10 days
CacheUnused http://www.ncsa.uiuc.edu/* 10 days
# ensure dynamically changing documents are only kept for
# short periods e.g. one modified 10 days ago will only last # 2 days
CacheLastModifiedFactor 0.2

# If a file was retrieved more than 5 days ago do a
# 'conditional get' request to the source server to check
# that it hasn't been updated in the meantime.
CacheRefreshInterval http://* 5 days
CacheRefreshInterval gopher://* 5 days
CacheRefreshInterval ftp://* 5 days

# CacheDefaultExpire ensures that Gopher and FTP files are
# cached. The default is 0 which is what we want for http
# documents with neither an expiry nor a last-modified stamp.
CacheDefaultExpire ftp://* 5 days
CacheDefaultExpire gopher://* 5 days

# Remove unwanted cached files at 3 am each day (garbage collection).
GeOn
GeDailyGe 3:00

Figure 7–4: A httpd.conf configuration file

Figure 7–4 shows a more complex configuration file. The various features are summarised below:

Parts 1 and 2 provides a mechanism for ensuring that the proxy gateway cannot be accessed from outside the local domain. Without these options it would be possible for a browser on an external system to use the proxy gateway to gain access to files which are restricted to local use.

Part 3 passes requests for the httpd, gopher, wais and ftp protocols.

Part 4 specifies the location for CGI files.

Part 5 is described above.

Part 6 describes the location and format of the server log file.

Part 7 specifies that server caching is to be available, and gives the location of the cache and the cache log files, together with the size (in Mbytes) of the cache.

Part 8 specifies the purging frequency for files in the cache.

Caching

Many clients provide client-side caching. This means that if you retrieve a file and then retrieve another file, when you return to the initial file, it will be retrieved from the client’s cache, thus saving a subsequent network transfer.

A number of servers also support caching by the server. This is illustrated in Figure 7–5.

Click here for Picture.

Figure 7–5: Caching by the server.

Caching can improve the performance of a WWW service by ensuring that frequently requested files will tend to be stored in the local cache. There is, of course, a danger that if the file on the remote server is updated that an out-of-date file will be retrieved from the cache. In practice, however, httpd server software which support caching can deal with this issue by, for example, looking at the date of the file on the remote server and, if the remote file is newer than the file in the cache, replacing the file in the cache with the new version of the file.

[Proxy Information]

http_proxy: www.leeds.ac.uk

gopher_proxy: www.leeds.ac.uk
Figure 7–6 Client Configuration File To Support Caching

It order for a client to make use of a cache, the client configuration file (e.g. the MOSAIC.INI file) must be suitably configured. Figure 7–6 illustrates the relevant options for the MOSAIC.INI file.

Caching Strategies

As well as using a local server cache, it is also possible to use a national caching service. The Unix HENSA service at the University of Kent at Canterbury run a national caching service. To use this service the local client should point to www.hensa.ac.uk.

An institution will need to decide whether to use a caching service and, if so, whether to have caching services running on a number of departmental system, to have an institutional caching service, or to use the national caching service at HENSA. In the future it may be possible to chain caches (this facility will provided in an experimental caching software known as Lagoon). The possibility in the long term of having institutional, metropolitan, national and continental caches should be considered.

Proxy Gateways

In many academic institutions off–campus access to the Internet is restricted to authorised computers. Depending on the institution’s local policy, authorisation may be restricted to computers located in offices in which there is an individual who is responsible for use of the machine. Such a policy may be enforced in order to provide some means of security against hacking remote services. However this policy would appear to prevent students from accessing remote information services from computers in open access cluster areas.

In practice there is a technique known as proxy gateways which can be used to provide access to services off–campus, without compromising local security. With a proxy gateway a trusted system (typically a Unix system which is more secure to hacking than a desktop machine) will have Internet access. Machines in open access cluster can point to the proxy gateway, which will then retrieve information from off–campus services.

It should be noted that with increasing usage of Internet services such as the World Wide Web, the author believes that the provision of security mechanisms, such as proxy gateways, will be increasingly important.

Further information

Further information on caching and proxies is available at the following URLs:

http://www.hensa.ac.uk/www94/

http://info.cern.ch/httpd/Proxies/ManyProxies.html

http://www.hensa.ac.uk/new.cache.html

Security

The httpd server also handles a number of security issues. It is common practice to restrict access to a certain area of the filesystem. For example if the server configuration files contains the lines:

map /* file:/apps/WWW/*

pass file:/apps/WWW/*

fail *

Figure 7–7 Server Configuration File.

then clients will only be able to access files held under the directory /apps/WWW/.

Note This statement refers to clients running on remote machines. If the client is running on the same machine as the server, the client will normally be able to access files on the server to which it has read access.

Additional levels of security can also be specified:

Restricting access to specified computer systems.

Restricting access by username and password.

The method of implementing such security tends to be server dependent, and will not be described in this document.

Which Server?
The most widely used servers are probably those developed at CERN and NCSA for the Unix platform. Unix is probably the best platform for running an institutional WWW service, since it is a mature, pre-emptive multi-tasking operating system. In addition, Unix provides a wide range of tools which can be used to assist in system administration. Servers are available for the PC and Macintosh platform, but, due to the inherent deficiencies in the operating system environments which are currently used on the platform, such servers are probably not recommended if you wish to run a large-scale, stable service.

Servers have been developed for the Windows NT environment. This may provide a robust operating system environment which can be used for providing a WWW server on an Intel platform.

Further Information

The document "Setting up a World Wide Web Server," which is available at the URL http://scholar.lib.vt.edu/reports/Servers–web.html, gives advice on setting up a server.

A collection of utilities intended especially for WWW system administrators is available at the URL ftp://src.brunel.ac.uk/WWW/managers/

8 Extending WWW

External Viewers

Access to WWW can be achieved by using a client such as NCSA Mosaic to display HTML documents and inline images in GIF format. However the World Wide Web is an extensible system: clients can access information which is in other formats than HTML.

When a client receives a file from a server it checks on the file type. If the file type indicates that it is an HTML document, the file will be displayed by the browser. Otherwise the browser's configuration file can specify an external viewer which can be used to display the file. A list of widely used external viewers is given in Table 8–1.

<table>
<thead>
<tr>
<th>File Format</th>
<th>Viewer</th>
</tr>
</thead>
<tbody>
<tr>
<td>JPEG</td>
<td>LVIEW (MS Windows) xv (X Windows)</td>
</tr>
<tr>
<td>Postscript</td>
<td>Ghostview</td>
</tr>
<tr>
<td>SVI</td>
<td>xdvi (X Windows)</td>
</tr>
<tr>
<td>MPEG</td>
<td>mpeg_play (X Windows and MS Windows)</td>
</tr>
</tbody>
</table>

Table 8–1 Popular Viewers.

The association between the file type and the viewer is given in the browser's configuration file. A typical configuration file for Mosaic for Windows is given in Figure 8–1.

[Vviewers]

TYPE0=audio/wav

TYPE1=application/postscript

TYPE2=image/gif

TYPE3=image/jpeg

TYPE4=video/mpeg

TYPE5=video/quicktime

TYPE6=video/msvideo

TYPE7=application/x-rtf

TYPE8=audio/x-midi

TYPE9=audio/basic

TYPE10=audio/x-action

TYPE11=application/x-w3launch

application/postscript="L:\winapps\ghostvysview %ls"

application/x-w3launch="m:\win32\onh\w3launch\w3launch %ls"

image/gif="L:\winapps\mosaic2\dview %ls"
Running Client Applications

If a Postscript file is retrieved from a WWW server the browser program normally responds "I don't know what to do with a Postscript file – but I know a program that does. I'll pass the Postscript file on to the Ghostview program". If, for example, an Excel spreadsheet is retrieved from a WWW server the client could be configured to respond "I don't know what to do with an Excel spreadsheet file – but I know a program that does. I'll pass the spreadsheet file on to the Excel program". This technique extends the functionality of the World Wide Web from acting as a distributed file viewer to acting as a distributed program manager.

Security Implications

Unfortunately there are a number of security loopholes in such an approach. An application developed using the Toolbook authoring system could be delivered using WWW. The application could then be launched using a local copy of Toolbook. The application could have a button marked Start. Clicking this button could then result in files held on the local machine being deleted! Even associating a word processed document with the Word For Windows holds dangers, as many Microsoft applications, including Word For Windows, support the use of macros, including autostart macros, which could also cause files to be deleted.

As a general principle there are dangers in automatically invoking applications from WWW clients.

Implementing Security – W3Launch

There are security problems in using a WWW browser to download and run software from the Internet. It is generally not considered wise to configure a browser so that it recognises file types which contain programs. Jon Maher, Biochemistry and Molecular Biology, University of Leeds has developed a launching program for the Bionet TIP project which provides a simple and secure method of launching only authorised software.

Further details on the W3Launch program is available at the URL
http://www.leeds.ac.uk/bionet/student/pre-stud.htm

Click here for Picture.

Figure 8-2 W3Launch

Server-side Extensions

Example

The previous section described how it is possible to run applications on the client machine. It is also possible to run software on the server. A simple application running on the server is shown in Figure 8.3

```sh
#!/bin/sh

echo Content-type: text/html

echo

if [ $# = 0 ]
then
  echo "<HEAD>"
  echo "<!-- Script written by Brian Kelly -->"
  echo "<TITLE>Search University Phone Directory</TITLE>"
```
echo "<ISINDEX>

echo "</HEAD>

echo "<BODY>

echo "<H1>Phone Directory</H1>

else

echo "<H1>Results of Search</H1>

fi

Figure 8–3 Application Running On The Server

The program, which is a C shell script which runs on the Unix server system, can be executed by selecting the URL http://www.leeds.ac.uk/ucs/cgi-bin/phone

When the URL is selected since no arguments are provided, the first part of the if statement is run. This will generate the following HTML document:

<HEAD>

<!-- Script written by Brian Kelly -->!

<TITLE>Search University Phone Directory</TITLE>

<ISINDEX>

</HEAD>

</BODY>

<H1>Phone Directory</H1>

Enter surname of the person you are searching for.<P>

Script written by <A HREF='http://www.leeds.ac.uk/ ucs/people/BKelly/bk.html'>Brian Kelly</A>

</BODY>

Figure 8–4 Application Running On The Server

The <ISINDEX> tag generates a search dialogue box. The HTML document is rendered as shown below:

Click here for Picture

Figure 8–5 Running The Script.
When an entry is entered in the Search box and then <Enter> key pressed, the script in Figure 8.4 is executed again. This time, since the program will be given an argument, the second part of the if statement will be executed. This will generate the HTML tags and then invoke the Unix grep command to search a file for lines containing the search string.

Click here for Picture.

**Figure 8-6 Output From The Script.**

**CGI Programs**

The example described above is known as a CGI program. CGI stands for the Common Gateway Interface. It is a standard which has been adopted by a number of server developers (primarily developers of the CERN and NCSA server software) for running programs on the server machine.

---

**9 Utilities**

A number of useful utility programs have been developed which will assist systems managers and information providers.

**w3new** is a program which will extract a list of URLs from the Mosaic client htlst file or extract URLs from a HTML document. It will then retrieve the modification dates for each document listed and output a HTML file with the URLs sorted by their last modification date.

Information about the program is available at the URL http://www.stuff.com/cgi-bin/bbcum?user=butter&pkg=w3new The utility was written by Brooks Cutter (mailto:butter@stuff.com).

**wusage** is a WWW server usage meter which produces weekly activity reports in HTML. In addition it provides graphical displays of server usage.

Further information is available at the URL http://siva.cshl.org/wusage.html The software is available from the URL ftp://isis.cshl.org/pub/wusage wusage was written by Thomas Boutell (mailto:boutell@netcom.com).

**getstats** (formerly called getaistes) is a versatile WWW server log analyser. It is available at the URL http://www.eit.com/software/getstats/getstats.html

**weblint** is a Unix utility for checking the syntax of HTML documents. The checks include illegally nested, overlapped, unclosed and obsolete tags. Further details are available at the URL http://www.khoros.umn.edu/staff/neilb/weblint.html The software can be obtained from the URL ftp://ftp.khoros.umn.edu/pub/perl/www/. The utility was written by Neil Bowers, Khoral Research Inc. (mailto:neilb@khoros.umn.edu) The email list weblint@khoros.umn.edu provides announcements of new versions of Weblint. Email Neil Bowers if you wish to be added to the list.

**Verify_links** is a robot which performs link verification. Further information is available at the URL http://wak.eit.com/wak/dist/doc/admin/webtest/verify_links.html

**MOMspider** (Multi-Owner Maintenance spider) is a tool which can be used to help information providers and system managers to maintain links to documents. MOMspider is available at the URL http://www.ics.uci.edu/WebSoft/MOMspider/

**Hypermail** is a program that converts a file of email message to a hypertext WWW form. It is available at the URL http://gunno.stanford.edu/html/hypermail/hypermail.html

The following utilities are available at the URL ftp://src.doc.ic.ac.uk/pub/packages/infoystems/WWW/tools

**checkweb** looks for dead links in your Web

**html+tables.shar** creates preformatted text tables from HTML + Table definitions

**mosaic-wais-clipl** do a WAIS search using Mosaic from the command line

**newstlist** compile an HTML page of links to all newsgroup on your server

**simon** URL database to replace NCSA Mosaic's Hotlist

**test-egi** sets up HTTP environment for a CGI script

**url-get.pl** a perl script which brings in any document given its URL

**w3get.pl** retrieves a HTML page named by a URL and all IMG's and IREF's in it
10 Legal and Ethical Issues

Is your WWW service legal? Who is legally responsible for the contents of a WWW service? Is pornography acceptable on a WWW service? If not, who defines what is pornographic and what is art? How do you reconcile control over the contents of a WWW server with intellectual freedom?

The author does not know the answer to these questions. Fortunately WWW is attracting the interests of lawyers, philosophers and artists who are starting to address these issues. Many of the papers which have been published address issues which affects WWW providers in the USA. The American Constitution, and in particular the amendment on free speech, means that much of the work published in the USA is not relevant to the UK.

A future edition of this handbook may contain information relevant to UK legislation and culture. In this edition a number of questions are raised which WWW service and information providers should consider.

Liability

It could be argued that the contents of a WWW service are the responsibility of the organisation which runs the service. So if an undergraduate has been granted to publish information and publishes libellous information the University may be legally responsible. An editorial in the Times Higher Education Supplement suggested that if the organisation has published guidelines covering acceptable and unacceptable use the organisation will have a strong defence if a case is brought to law.

Pornography

Are pictures of naked women acceptable on a WWW service? It could be argued that similar guidelines which govern the contents of a University library should be developed for the WWW. Are pictures of naked women acceptable in books in the university library? The answer is probably yes, especially if the university has a fine art department. Similar arguments could be made for textual pornography.

However even the most liberal individual is likely to be offended by some of the pornography which is believed to be available on the Internet. In addition UK legislation on computer pornography is likely to be introduced shortly. This could mean that universities have a legal obligation to concern itself with computer pornography.

Copyright

A WWW manager may have a responsibility to ensure that copyright material is not made available unless the copyright holder has granted permission. This may affect research papers which have been submitted for publication. It may also affect the use of photographs, for which the copyright may be owned by the photographer or the organisation which commissioned the photograph.

Data Protection Act

Information about individuals which is available on WWW may have to be registered with the Data Protection Officer. The information provider may have to abide by regulations to ensure the accuracy of the information.

Equality Of Access To Information

WWW can provide global access to a wide range of information services. However including large logos and graphical icons on pages can act as a barrier to access to the information, especially for readers in developing countries will limited network access. In some developing countries access may be provided over local telephones lines. A health worker in a hospital in Africa who wishes to retrieve information about public health services may have to pay the additional in retrieving unnecessary graphics. If the local telephone company is owned by a multinational telephone corporation then accessing the information will result in a transfer of money from the developing country to the multinational corporation.

Advertising

As shown in Figure 10–1 some WWW service providers have sponsors for their pages. Is this currently acceptable within the UK academic community? Should it be acceptable?

Click here for Picture.

Figure 10–1 The What's New On Mosaic Page

What Is Your WWW Service For?

Formulating an institutional acceptable use policy for WWW information providers may not be a simple task. There are likely to be lively discussions over censorship and control. The formulation of the policy will be helped if the institution has a clear idea of what it expects from its WWW service. Is it:

To provide easy access to high quality information.
To enable students to gain experience in multimedia publishing.
To provide researchers with a test bed for their research.
To provide members of the University with access to cultural services and to enable them to publish cultural services.

Further Information
An interactive document called Sex, Censorship and the Internet is available at the URL http://www.chn.org.80/CAF/cafuic.html. This document asks questions such as should universities carry alt.sex Usenet groups and should students be punished for using vulgarities on the Net. The document provides pointers to case studies.

Cranfield University have published guidelines for information providers which is available at the URL http://www.cranfield.ac.uk/docs/publish_code.html

11 CWIS & WWW

WWW is an ideal system for developing a campus (or community) wide information system (CWIS). The world’s first multimedia CWIS was developed at the Honolulu Community College and officially announced at the end of May 1993. It is available at the URL http://www.hcc.hawaii.edu/

Click here for Picture

Figure 11–1 CWIS At HCC

The HCC CWIS was developed to support its goal of becoming the “Technological Training Centre of the Pacific”. The most important aspects of developing and managing an effective CWIS are managerial and not technical. Formulating the objectives of a CWIS, resourcing it and developing a training programme are key issues which an institutional needs to address.

The Universities and Colleges Teaching, Learning and Information Group (UCTLIG) are producing a CWIS Manager’s Handbook which will address many of these issues.

Find Out More

Papers by Judy Hallman about CWISes are available at the URL ftp://sunsite.unc.edu/pub/docs/about-the-net/cwis/cwis-1 and ftp://sunsite.unc.edu/pub/docs/about-the-net/cwis/hallman.txt

Polly–Alida Farrington’s listing of CWISes is available at the URL http://www.rpi.edu/internet/cwis.html

Lists of (global) CWISes are available at the URLs http://www.rpi.edu/Internet/cwis.html and http://kawika.hcc.hawaii.edu/ws94/cwis.html

The CWIS–L Listserv mailing list provides a forum for the discussion of topics related to campus–wide information systems. To subscribe send the message SUB CWIS–L your name to the address LISTSERV@MSU.EDU

12 Teaching And Learning On WWW

Although WWW was initially used as a distributed multimedia system techniques such as CGI scripts meant that interaction could be built into WWW applications. Much of the interest in the WWW within the academic community is based on its potential for developing distributed teaching and learning software rather than simply delivering information.

Examples of Teaching And Learning On WWW

An early example of a distributed multimedia teaching prototype was developed by Ben Whitaker, School of Chemistry, University of Leeds in 1993. As can be seen in Figure 12–1 this prototype is a simple hypertext application. It is of interest because it illustrates how distributed teaching applications can be developed.

Click here for Picture

Figure 12–1 Early Example Of A Distributed Multimedia Teaching Application.

A more sophisticated teaching application was developed by the School of Chemistry in conjunction with Imperial College. The example illustrated in Figure 12–2 makes use of a chemistry MIME type.
Figure 12-2 Using a MIME Chemistry Type.

In this example the WWW client is configured to associate the MIME type with the RasMol program. For example in NCSA Mosaic for X the line:

```
chemical/x-pdb; rasmol %s
```

is included in the .mailcap file. When a URL with the extension .pdb is selected the file will be downloaded and the RasMol program launched, as illustrated in Figure 12-2.

Further information on this project is available at the URL http://chem.leeds.ac.uk/Project/MIME.html

The Globewide Network Academy (GNA) is a consortium of educational and research organisation. Its mission is to provide a central organisation in which students, teachers, scholars and researchers can meet and interact. Further information about GNA is available at the URL http://uu-gna.mit.edu:8001/uu-gna/

Mark Cox, Department of Industrial Technology, University of Bradford presented a paper at the Mosaic and the Web conference on Robotic Telescopes: An Interactive Exhibit on the Web. This paper is available at the URL http://www.eia.brad.ac.uk/mark/www94/www94.html

Mark also has a collection of pointers to hardware control services over the Web which is available at the URL http://www.eia.brad.ac.uk/mark/fave-inter.html

A Virtual Frog Dissection Kit has been developed at the LBL. It is available at the URL http://george.lbl.gov/ITG/html/docs/WholeFrog/WholeFrog.html

Click here for Picture

Figure 12-1 Frog Dissection.

CD ROM Facilities

Providing teaching and learning services on WWW does not necessarily deny access to those who do not have a network connection. Teaching and learning services developed on WWW can be transferred to a CD ROM and used on a standalone system. Such systems are typically developed so that there is a closed set of links. The files (which could include HTML documents, image, sound and video files) and the WWW browser software can then be transferred onto a CD ROM. This approach provides an updateable service for users with network connectivity together with a fixed service for users with access to a PC or Macintosh with a CD ROM player.

National Resources

A number of CTI (Computers in Teaching Initiative) and TLTP (Teaching and Learning Technology Programme) projects are using WWW to disseminate information about their projects or, in some cases, to deliver their courseware.

CTISS is available at the URL http://www.ox.ac.uk/ctiss/

CTI Centre For Biology is available at the URL http://www.liv.ac.uk/ctibiol.html

CTI Centre For Chemistry is available at the URL http://frs2.liv.ac.uk/ctichem.html

CTI Centre For Law is available at the URL http://croesus.csv.warwick.ac.uk/WWW/law/default.html

CTI Centre For Psychology is available at the URL http://ctipsych.york.ac.uk/

CTI Centre For Sociology is available at the URL http://lorne.stir.ac.uk/Departments/cti_centre/

CTI Centre For Textual Studies is available at the URL http://www.ox.ac.uk/depts/humanities/

BioNet Project is available at the URL http://www.leeds.ac.uk/bionet.html

CLIVE Project is available at the URL http://www.vet.ed.ac.uk/.

Insurrect Project is available at the URL http://av.avc.ucl.ac.uk/

Institute Of Computer Based Learning, Heriot-Watt is available at the URL http://www.icbl.hw.ac.uk/

INTERACT Project is available at the URL http://medusa.eng.cam.ac.uk/~interact/

Interactive Learning Centre, University of Southampton is available at the URL http://ilc.ecs.soton.ac.uk/welcome.html
13 Collaboration On WWW

WWW was originally envisaged by Tim Berners-Lee as a groupware tool. In practice it grew in popularity as a publishing tool. However software developers are now working on tools which will facilitate collaboration on WWW. A brief summary of some of the collaborative tools is given below.

Asynchronous Systems

WIT

WIT, the WWW Interactive Talk system, was announced shortly after the WWW 94 conference in CERN. WIT can be accessed at the URL http://info.cern.ch/wit

Click here for Picture.
Figure 13–1 WIT.

Access To Usenet

A number of WWW browsers, including Netscape, provide the ability to post to Usenet newsgroups.

Click here for Picture.
Figure 13–2 Posting To Usenet News.

Hypermail

Hypermail is a utility which can be used to convert mail archives to hypertext format on WWW. A example of a hypermail archive is illustrated below.

Click here for Picture.
Figure 13–3 A Hypermail Archive.

Mailserv

Mailserv provides a forms interface to a number of mailing list servers. The software is available at the URL http://hquest.com/~fitz/www/mailserv/ The software was written by Patrick M Fitzgerald (mailto: pmfitzge@hquest.com)

Click here for Picture.
Figure 13–4 The Mailserv Interface To Mailing List Servers.

Synchronous Systems

Video conferencing facilities are being developed which can be integrated with WWW.

Click here for Picture.
Figure 13–5 Accessing A Video On WWW.

One interesting application of a multimedia desktop conferencing systems is MONET (Meeting on the Network) which is described in Applications of Mosaic in Health Care Delivery by Srivasa et al. This paper, which was presented at the Mosaic and The Web conference, is available at the URL.
http://www.ncsa.uiuc.edu/SDG/IF94/Proceedings/ModTrack/srivasa/artemis.html

Click here for Picture.

Figure 13–6 Accessing A Video On WWW.

At the time of writing many of these services are experimental. However, given the rapid growth of WWW and the extent of development work which is going on, such services may be mainstream in the near future.

Virtual Conferences

One form of collaboration within the academic community is through conferences, workshops and seminars. Whenever the author gives a paper at a conference or is involved in running a workshop or a course he makes his papers, OHP foils, etc. available on WWW under his personal page (sometimes referred to as a vanity page).

About 200 of the papers which were given at the second WWW conference, Mosaic and The Web, were available on WWW before the conference began. Perhaps one important question which the academic community should be addressing is whether it should be the standard practice for conference proceedings to be made available on WWW.

Further Information

A collection of WWW collaborative projects is available at the URL
http://union.ncsa.uiuc.edu/~liberte/hypernews/related.html

A list of conference proceedings available on WWW is given in Appendix 7.

14 Libraries & WWW

University Libraries should have a strong interest in WWW developments. This handbook provides a overview of the World Wide Web which should be of interest to libraries which are considering using WWW.

Example Of A Gateway To A Library Catalogue

In the UK many university library catalogues are held in proprietary systems with old–fashioned user interfaces. It may be possible, however, to use WWW to provide an interface to the library catalogue which is consistent with other information services on WWW. At the University of Leeds a backup copy of the library catalogue is kept on a central Unix system using the BRS free text retrieval system. A gateway program has been developed by the Computing Service which provides access to the Library catalogue. This gateway is available at the URL
http://www.leeds.ac.uk/library/cats/backup.html

Click here for Picture.

Figure 14–1 Gateway To A Library Catalogue.

In Figure 14–1 a form is completed. The term Internet is used as a search term. Once the form is submitted the data is sent to a CGI program. In this case the CGI program is a C program which invokes the BRS free text retrieval system. The output from the BRS program is then processed to generate the appropriate HTML markup. The output from the search is illustrated in Figure 14–2.

Click here for Picture.

Figure 14–2 Gateway To A Library Catalogue.

Resources

The Clearinghouse for subject–oriented Internet resource guides is available at the URL
http://http2.slis.umich.edu/~lou/chhome.html

The EINet Galaxy collection of online resources is available at http://galaxy.einet.net/galaxy.html

The CERN Virtual Library is available at the URL
http://info.cern.ch/hypertext/DataSources/bySubject/Overview.html

The Boulder Community Network service is available at the URL http://bcn.boulder.co.us/ Its policy statement is available at the URL http://bcn.boulder.co.us/bcn/policy.html The policy statement includes a bill of rights, a freedom to read statement and a freedom to view statement.

The following Library resources may also prove useful:
http://www.lib.washington.edu/~tdowling/libweb.html
Finding Out More

Web4Lib is a mailing list aimed at library-based WWW managers and developers. To subscribe to the list email listserv@library.berkeley.edu with the message SUBSCRIBE Web4Lib yourname.

16 Future Developments

This handbook describes how to run a WWW service using the technology which is available today. However the technology is developing so rapidly that it is important that WWW managers and information providers are aware of developments which may happen sooner rather than later.

Uniform Resource Identifiers

Uniform Resource Locators (URLs) describe the location of a resource on WWW and the protocol which is used to access the resource. An object on WWW may be available in many locations: for example popular browsers, such as NCSD Mosaic, are available from anonymous FTP servers in many locations around the world. The mirroring of files helps to minimise network traffic over busy links, such as the trans-Atlantic link. Mirroring also reduces the load on the central server. Uniform Resource Names (URNs) will provide a mechanism for uniquely identifying a resource. In the future it is likely that a browser will request a URN rather than a URL. A URN to URL resolver will locate the nearest object (nearest in network terms).

Uniform Resource Characteristics (URCs) will provide meta-information about a document. This information could include keywords, expiry dates (for caching servers), copyright and cost information. URCs could also provide information about the quality of the document. For example a seal of approval (SOAP) could be given by university publications group which confirms, by the use of a digital signature, that the document is a PhD thesis.

Uniform Resource Identifiers (URIs) includes URLs, URNs and URCs. The URI specification is available as RFC 1630. The mailing list uri@bunyip.com is used to discuss URNs. Send email to uri-request@bunyip.com to subscribe to this list. Archives of the list are available at the URL http://www.acrlanl.gov/URI/archive/uri~archive.index.html

New Facilities

CCI

NCSD Mosaic For X (version 2.5) provides support for CCI (Common Client Interface). This will provide a standard mechanism by which WWW browsers can communicate with external programs. A number of demonstrations of this facility are available, including a slideshow program, which instructs Mosaic to display URLs which are specified in a file. A program called xwebtech provides a mechanism by which a teacher can control the display of Mosaic on student's machines.

W3A

W3A (World Wide Web Applets) is a proposal for a standard API for dynamically linking applets (which can be defined as a piece of software that can be attached to a host program such as a WWW browser).

Appendix 1 Mailing Lists

This section contains information on mailing lists and Usenet groups on topics related to the World Wide Web.

Please note that before sending a message to any of these lists please listen to the discussions first and, where possible, read the information about the list. You should not send simple questions about, say, installing Mosaic on your home PC to a list for developers of the WWW protocols.

Usenet News

cmp.infosystems.www.users

cmp.infosystems.www.users provides a forum for discussion of WWW client software (such as Mosaic, Cello and
Lynx), new user questions, client setup questions, client bug reports, resource discovery questions on how to locate information on WWW that can't be found by the FAQ and comparisons between various client packages are among the acceptable topics for this group.

**comp.infosystems.www.providers**

*comp.infosystems.www.providers* provides a forum for the discussion of WWW server software and the use of server software to provide information to users. General server design, setup questions, server bug reports, security issues, HTML page design and other concerns of information providers are among the likely topics for this group.

**comp.infosystems.www.misc**

*comp.infosystems.www.misc* provides a general forum for discussing WWW issues which are not covered by the other comp.infosystems.www groups.

**comp.infosystems.announce**

*comp.infosystems.announce* is for announcement of new information services (e.g. new WWW sites) and new software products (new server software, new clients, new document converters, etc.)

**comp.infosystems.wais**

*comp.infosystems.wais* covers WAIS topics, including integration of WAIS with WWW.

**comp.text.sgml**

*comp.text.sgml* covers SGML, including HTML.

**comp.infosystems.announce**

An archive of the comp.infosystems.announce Usenet group is available at the URL http://www.cs.rochester.edu/users/grads/ferguson/announce/

Archives of These Groups

Archives of the **www-announce, net-happenings** mailing lists and *comp.infosystems.www.* Usenet groups are available at the URL http://cair-archive.kaist.ac.kr/Archive/Announce/

**CERN Mailing Lists**

To join a list at CERN send electronic mail to listserv@info.cern.ch with the line subscribe www-list your name.

For example if John Smith wanted to subscribe to the *www-announce* list he would send the message subscribe www-announce John Smith

An overview of CERN mailing lists is available at the URL http://info.cern.ch/hypertext/WWW/Administration/Mailing/Overview.html Alternatively send an email message to listserv@info.cern.ch containing line lists to receive a list of lists or review list to receive a list of subscribers to list.

www-announce

For anyone interested in WWW, its progress, new data sources, new software releases. Please refrain from postingadministrivia to this list! The list owners want to keep it low volume, large membership.

**www-html**

Technical discussions of the HyperText Markup Language HTML and HTML Plus designs. Design discussions only, please, not newcomer questions.

This list is archived at the URL http://198.92.133.3/menus/6581.htm and at the URL http://gummo.stanford.edu/html/hypermail/hypermail.html

**www-talk**

Technical discussion for those developing WWW software or with that deep an interest. (Please keep this to WWW technical design only. *Not* general questions from non-developers, which should go to the newsgroup, nor for HTML topics which should go to www-html.)

This list is archived. A threaded version of the archive is available at the URL http://gummo.stanford.edu/html/hypermail/archives.html.

Another archive is available at MCCMedia's archive at the URL http://www.verity.com/menus/6581.htm
www-rdb
Discussion of gatewaysing relational databases into WWW. It is archived at the URL http://info.cern.ch/hypertext/WWW/Archive/www-rdb

www-proxy
Technical discussion about WWW proxies, caching, and future directions.
It is archived at the URL http://info.cern.ch/hypertext/WWW/Archive/www-proxy

Other Mailing Lists

web-support
A Mailbase list which can be used for discussions about WWW issues. Based in the UK. To subscribe email mailbase@mailbase.ac.uk with the message join web-support yourname

The web-support archives are available at the URL http://mailbase.ac.uk/pub/lists/web-support

unite
A Mailbase list which can be used for discussions about a User Network Interface To Everything. Based in the UK with an international membership. mailbase@mailbase.ac.uk with the message join unite yourname

The UNITE archives are available at the URL http://mailbase.ac.uk/pub/lists/unite

www-managers
The aim of this list is to provide a high signal-to-noise, quick turn-around forum for managers of WWW servers and sites to get answers to specific questions about the setup and maintenance of HTTP servers and clients. The mailing list is managed by a utility called majordomo. To subscribe send the message subscribe www-managers to the address majordomo@lists.stanford.edu

Mosaic-I.
A Listserv list for the NCSA Mosaic WWW browser. To subscribe send the message subscribe mosaic-I firstname lastname to the address listserv@uicvm.uic.edu

NOTE This list is now believed to be defunct since it was being used for basic Mosaic questions, rather than providing a forum for Mosaic developers.

Netscape
A Listserv list for the Netscape WWW browser. This list is for the purpose of discussing features and bugs contained in this new browser, as well as the new tags Netscape implements. To subscribe send the message subscribe netscape firstname lastname to the address listserv@irlearn.ucd.ie

HTML Working Group
A mailing list for an IETF working group which is discussing developments of HTML. To subscribe email html-wg-request@oel.org with the message SUBSCRIBE html-wg yourname An archive of the list is available at the URL http://www.ics.uci.edu/pub/ietf/html/

Web4lib
A list for Library-based WWW managers and developers. To subscribe email listserv@library.berkeley.edu with the message SUBSCRIBE Web4lib yourname

VRML
Virtual Reality Markup Language or VRML is an evolving specification for a platform-independent definition of 3-dimensional spaces within the World Wide Web. It is designed to combine the best features of virtual reality, networked visualization, and the global hypermedia environment of the World Wide Web.

To subscribe to the Virtual Reality Markup Language (VRML) list send mail to majordomo@wired.com containing the message subscribe www-vrml

Further information is available at the URL http://www.wired.com/vrml/

Quality
A mailing list for the discussion of quality issues. To subscribe to the list send the message subscribe quality to
the address listmanager@naic.nasa.gov.

An archive is available at the URL http://naic.nasa.gov/naic/archives

MacHTTP talk

A mailing list for MacHTTP users has been set up. It provides an open forum for any questions, answers, suggestions, announcements, etc. about the MacHTTP server software. To subscribe to the list send a mail message to the address listserv@oac3.hsc.uth.tmc.edu containing the message subscribe machhttp_talk
first name surname

Further information is available at the URL http://www.uth.tmc.edu/mac_info/machhttp/mailing_list.html

cello-l

Cello-L is a discussion list for users of the Cello WWW browsers. To subscribe to the list send the message sub cello-l your name to the address listserv@cornell.edu Further information is available at the URL http://ftp.law.cornell.edu/pub/LII/Cello/default.htm The Cello FAQ is available at the URL Archives of the list are available at the URL gopher://gopher.law.cornell.edu:70/11/listserv/cello

libwww-perl

libwww-perl is a library of Perl4 packages which provides a simple and consistent programming interface to the World-Wide Web. This library is being developed as a collaborative effort to assist the further development of useful WWW clients and tools.

A mailing list has been established for technical discussion about libwww-perl, including problem reports, interim fixes, suggestions for features, and contributions. The mailing list address is libwww-perl@ics.uci.edu and administer (including subscribe requests) should be sent to libwww-perl-request@ics.uci.edu

A Hypermail archive of the mailing list is also available at the URL http://www.ics.uci.edu/WebSoft/libwww-perl/archive/

moo-www

moo-www is a mailing list to discuss links between MUDS, in particular systems based on Pavel Curtis’s MOO server, and the World Wide Web. Subjects for discussion include:

In-MOO clients to access the web.

Web clients that understand how to talk to MUDs.

Using the Web to provide Multimedia capabilities to MUDs.

The list is at moo-www@maths.tcd.ie Subscription requests should go to moo-www-request@maths.tcd.ie

www-speed

www-speed list is dedicated to the proposition that the web is just too darned slow, and that some of its key components have inherent performance problems that cannot be dealt with without changes to protocols. Topics appropriate to the list are:

Performance issues in existing implementations

Managing busy servers

Benchmarking servers

High performance Server design

High speed (130Mbps+) issues

Client implementation

HTTP - beyond 1.0

Other performance and scalability issues

The list address is www-speed@tipper.oit.unc.edu The request address is www-speed-request@tipper.oit.unc.edu

WWW Courseware List

This is a list dedicated to courseware on WWW. To subscribe send mail to www-courseware-request@eit.com
containing the message subscribe

An archive of the list is held at the URL http://www.eit.com/mailinglists/www-coursesware/archive/

WWW Literature List

This is a list dedicated to literature on the WWW. To subscribe send mail to www-literture-request@eit.com containing the message subscribe

An archive of the list is held at the URL http://www.eit.com/mailinglists/www-literture/archive/

WWW Security List

This is a list to discuss different methods of providing a secure WWW service. The list will focus on how to secure HTTP and/or HTTP-like protocols to provide privacy, user authentication, service certifications and document checking (digital signatures).

To subscribe send mail to www-security-request@nsx.rutgers.edu containing the message subscribe www-security

An archive of the list is held at the URL http://www.verity.com/www-security.html

Information about the www-security list is also available at the URL http://www.ns.rutgers.edu/www-security/index.html

WWWBuy.info

Discussions of issues of commercial transactions of information via the Web take place on the www-buyinfo mailing list. To subscribe send the message subscribe www-buyinfo to the address www-buyinfo-request@allegra.att.com

The archives are held at the URL http://www.research.att.com/www-buyinfo/about.html

Appendix 2 WWW Resources

A wide range of resource materials about the World Wide Web are available on the World Wide Web. A number are listed below.

WWW Online Resources

Cyberweb is available at the URL http://www.charm.net/~web/

Spider's Web is available at the URL http://gagme.wwa.com/~boba/spider.html

One World is available at the URL http://oneworld.wa.com/htmldev/devpage/dev--page1.html

Web Weaver's Page is available at the URL http://www.nas.nasa.gov/RNR/Education/weavers.html

WebStars: Astrophysics in Cyberspace is available at the URL http://guinan.gsf.nasa.gov/

Pointer's to WWW resources (Toronto University) is available at the URL http://www.atirc.utoronto.ca/

PC Week's pointers to WWW resources is available at the URL http://www.upcweek.ziff.com/~pcweek/pointers.html

OsloNet is available at the URL http://www.oslonet.no/html/demo/WWWinhtml


WWW Icons and Clip Art

A list of online resources of icons and clip art which can be used to produce HTML documents containing graphics is given below. Note, however, that before using graphics in HTML documents you should be aware of the additional loads which will be placed on network and servers.


http://white.nosc.mil/images.html

http://guinan.gsf.nasa.gov/Alan/Richmond.html
WWW Conferences

Conference proceedings from the first WWW conference, WWW '94, held at CERN on 25–27 May 1994 are available at the URL http://www.elsevier.nl/

Further information about the second WWW conference Mosaic and The Web, held at Chicago on 17–20 October 1994 is available at the URL http://www.nccs.unc.edu/SDG/IT94/IT94Info.html A searchable index of the papers is available at the URL http://www.verity.com/spidersearch.html

The third WWW conference will be held at Darmstadt, Germany on 10–14 April 1995. Further details are available at the URL http://www.idg.fhg.de/www95.html

Other Resources

WWW Information At CERN

Information about the World Wide Web Initiative is available at the URL http://info.cern.ch/hypertext/WWW/TheProject.html

Best of the Web

The Best of the Web awards promotes WWW to new and potential users and helps information providers by demonstrating what can be done on WWW. The award winners and entrants are available at the URL http://wings.buffalo.edu/content/

WWW FAQ

The WWW Frequently Asked Questions (FAQ) is available at the URL http://sunsite.unc.edu/boutell/faq/www_faq.html

WWW FAQ


Information Superhighway in the UK

Information about the Information Superhighway in the UK is available at the URL http://tin.ssc.plym.ac.uk/isp.html

Appendix 3 National UK Services

Services

The Bulletin Board For Libraries (BUBL) holds a wide range of information of interest to anyone involved with libraries in education. Further information is available at the URL http://www.bubl.bath.ac.uk/BUBL/home.html

The Mailbase mailing list service run a WWW server which is available at the URL http://mailbase.ac.uk/

The Micros Hensa service run a WWW server which is available at the URL http://micros.hensa.ac.uk/

The Unix Hensa service run a WWW server which is available at the URL http://unix.hensa.ac.uk/

CTISS run a WWW server which is available at the URL http://www.ox.ac.uk/ctiss/

The Office for Library and Information Networking (UKOLN) runs a WWW server which is available at the URL http://ukoln.bath.ac.uk/UKOLN/home.html

CCTA, the UK Government computer agency, runs a WWW server which is available at the URL http://www.open.gov.uk/

NISS is setting up a WWW server which is available at the URL http://www.niss.ac.uk/ The server is due to be released by January 1995.

A TLTP specific Web Server is available at the URL http://www.icbl.hw.ac.uk/tltp
The Social Sciences Information Gateway is available at the URL http://sosig.esrc.bris.ac.uk/

Directories

A UK tourist guide is available at the URL http://www.cs.ucl.ac.uk/misc/uk/intro.html

A UK sensitive map is available at the URL http://scitsc.wlv.ac.uk/ukinfo/uk.map.html This service is maintained by the School of Computing and Information Technology, University of Wolverhampton (email jphb@scitse.wlv.ac.uk)

WAIS Resources

The following WAIS services are provided by NISS.

NISS Bulletin Board

A wide range of information of interest to varying sectors of the academic community. This service is available at the URL wais://gopher.niss.ac.uk/NISSBB

World Factbook

Basic details (population, climate, main industries and so on) for the countries in the World. Use a search term such as the name of a country to locate particular records. This service is available at the URL wais://wais.niss.ac.uk/World_Factbook

Roget's Thesaurus

The 1911 edition (enhanced with an additional 1,000+ words not included in the original version) of the ever-useful thesaurus of the English language. Use any word as your search term. This service is available at the URL wais://wais.niss.ac.uk/Roget

JANET News

JANET News contains material about the JANET computer network, such as registered domain names and addresses, and information about gateways to other networks. This service is available at the URL wais://news.janet.ac.uk/JANET.news

CHEST Directory

The CHEST Directory of software is available at the URL wais://wais.niss.ac.uk/CHEST_Directory

Appendix 4 References

Books


"Teach Yourself HTML Web Publishing in a Week" by Laura Lemay, to be published by Sams' Publishing (ISBN 0-672-30667-0). This book discusses not only the various aspects of HTML, Web servers, gateways, forms, and imagemap, but also focuses strongly on style and structure and navigation. In other words, its not just a reference guide. It also a style guide.

"Authoring HTML for Fun and Profit" or "HTML For Fun and Profit" by Mary Morris, to be published by Prentice-Hall. It includes forms, clickable images, server includes, indexing, linking and basic formatting. It will have a CD-ROM with examples and tools on it.


Magazines


Many magazines are being published which cover various aspects of the Internet. The following list gives some of the main ones, including ones published in the UK.

.net published by Fortune Publishing Ltd. Further details are available at the URL http://www.futureren.com/home.html or by sending email to netmag@futurenet.co.uk

infoHighway ISSN 1355–2465. For further details send email to p.deacon@eurodollar.co.uk or david@pipex.net

Wired. Further details are available at the URL http://www.hotwired.com/ For subscriptions details send email to subscriptions@wired.com

3W cost [$sterling]24 for 6 issues. Further information is available at the URL http://www.3w.com/3W/

---

Appendix 5 Conferences On WWW

Bruce Altner (mailto: ari@clark.net), the Director of Technical Services of ARInterNet Corporation has a vision for gatherings at the electronic meeting hall combines the best features of the WWW (browsing, multimedia and hypertext capabilities, searching and information retrieval, file downloading and e-mail communication, to name just a few) within the format of the traditional poster paper session.

Electronic Conferences and Workshops

Here are some real life examples of Electronic Conferences and Workshops:

ChemConf'93 is available at the URL gopher://info.umd.edu:901/1/inforM/Educational_Resources/Faculty_Resources_and_support/ChemConference

NASA High Alpha Conference IV (high angle of attack) is available at the URL http://www.dfrr.nasa.gov/Workshop/HighAlphaIV/highalpha.html

The HIDEC Electronic Conference (the F–15 Highly Integrated Digital Electronic Control program) is available at the URL http://mosaic.dfrr.nasa.gov/Workshop/HIDEC/Conf.DIRS/htmllinks/ConfWeb.html


...and its post–conference workshops is available at the URL http://xxx.lanl.gov/cmp-lg/ACL–94-post.html

1st Electronic Conference in Computational Chemistry (ECCC) is available at the URL http://hackberry.chem.niu.edu:7000/ECCCinformation.html

Reviews of Electronic Conferences

A thorough discussion of the pros and cons of this type of online gathering, see the review by the ChemConf'93 organizer Dr. Tom O'Haver which is available at the URL gopher://info.umd.edu:901/00/inforM/Educational_Resources/Faculty_Resources_and_support/ChemConference/BackgroundReading/OnlineConferencing.txt

And as a wonderful example of self–referencing, a la Douglas Hofstadter's Godel, Escher, and Bach, see the URL http://www.automatrix.com/conferences

An example of an "after–the–fact" online conference is available at the URL http://stardust.jpl.nasa.gov/igarsig/igarsig.htm
The scramble to save Web history

Obsolete equipment poses challenges for Internet historians

By Jeanene Harlick
STAFF WRITER

Old Web pages don't die, they just fade away.

In fact, web pages are blinking away so fast that the very history of these pioneer Internet years is threatened with extinction.

That's why a small group of scientists and students is racing to find ways to archive web pages, just as paper documents like the original Declaration of Independence have been preserved.

The difference, said archivist Jean Dekken, is that paper persists and, so far, objects in the digital world don't.

"They don't give any warning they're going to corrupt. They simply very quietly pass out of usefulness," said Dekken, who is with the Stanford Linear Accelerator Center.

A few years ago she found herself trying to electronically chase down the nation's first Web page, which was born in 1991 in a SLAC physicist's office.

No record of the page existed. Plowing through old backup tapes, the center's Web master found it, a page of plain text that made virtual history.

If the pair had waited one year, the page would have been lost forever, trapped in a file created by obsolete software no current hard drive can read.

It's too late, however, for the first e-mail ever sent. Too late, as well, for countless original World Wide Web pages and other documents stored electronically.

Confronted with voluminous electronic records, storage media that is obsolete within a few years and employees who delete documents before anyone realizes they're important, some fear future historians will look back on this time as a "digital dark ages."

"The chronicle of our entire period is in jeopardy," computer scientist Jeff Rothenberg wrote in a 1995 article that sent a shock wave through the digital community. "It is only slightly facetious to say that digital information lasts forever -- or five years, whichever comes faster."

Information overload
A look at the National Archives and Records Administration shows just one example of the depth of the problem.

Over the past quarter-century, the National Archives has been able to process just 100,000 e-records. But the Treasury Department alone is generating nearly one million e-mail messages a year that should be permanently preserved, U.S. Archivist John Carlin estimates.

In 1976, when the National Archives and Records Administration went looking for the 1960 U.S. Census data, officials realized it was stored on UNIVAC tapes whose tape drives -- the first main frame computers -- no longer existed. It took three years and some engineering back flips to eventually recover the data.

In addition, the National Archives has no digital archives or guidelines on how federal agencies should store electronic records the agencies are mandated to preserve.

"New technological ways to create records have not been matched by technological advancements to manage them," Carlin told Congress in October. "Unless we successfully address the key issues, essential evidence will be lost."

**Lost in the ether**

The first challenge posed by digital documents is one of physical presence. Many exist only in electronic format. They are e-mails and Web pages deleted and revised before a permanent copy is retained.

"These really are incredibly fugitive archives," Deken said. "They disappear. They disappear unless somebody takes appropriate steps early enough."

Archiving has been particularly hard for the Linear Accelerator Center because the center records most of its work on internal Web servers accessible to physicists worldwide. Researchers often update individual Web pages without saving the first version.

Because the Linear Accelerator Center is funded by the Department of Energy, it is required to send National Archives documentation of big projects like GLAST, the gamma-ray telescope the lab is now building.

Currently, Deken and Web master Joan Winters deal with the problem by sifting through backup tapes -- which take periodic snapshots of the center's Web pages -- every year to find important records. It's a time-consuming and inefficient process, and one Deken hopes will soon change.

Deken said organizations like the Linear Accelerator Center need a computer program that somehow automatically archives important Web documents as soon as they're created.

**Change, change, change**

The second challenge is the instability of electronic media.

Floppy disks are easily erased by stray magnetic fields or simple material decay.

But the more vexing problem is that hardware used to create documents and software used to interpret them are obsolete within five years, according to a National Media Lab study. Archivists are left with a disk -- or CD, or whatever future records are stored on -- full of ones and zeros as impossible to translate as hieroglyphics before the Rosetta Stone.

For example, eight-inch floppy disks -- the standard storage media for e-files in the 1970s -- are now unreadable.

It's like trying to find a place to play that old eight-track of Willie Nelson's greatest hits.
At the Linear Accelerator Center, they created their first Web pages using an operating system called IBM Virtual Machine (VM). Five years later, when Web master Joan Winters went looking for the pages, VM was being phased out and replaced by Unix. If Winters had waited until October 1998, when the center's VM machines were chucked, the pages would have been trapped in a virtual no man's land.

**What's the solution**

The Internet Archive, a nonprofit organization in San Francisco, started recording the World Wide Web in 1996. Using Web crawlers that automatically collect pages from public servers, the organization gets a complete snapshot of the Web every two months. So far, 14 terabytes, or one billion pages, have been collected.

It could be the world's first virtual museum, but in this case it's co-location facilities instead of climate controlled rooms that are preserving historical documents. Fort Knox-like buildings full of humming servers, the facilities are located throughout the Bay Area.

"We took a look around and we could see that the Internet, and the Web in particular, was full of an enormous amount of valuable scientific and cultural data," said Marliita Kahn, managing director. "It disappears almost as quickly as it appears. We didn't want it to go the route of television," for which no archives exist.

Historians, researchers and scholars can access the archives for free. The Xerox Palo Alto Research Center recently used the snapshots to study Web surfing behavior and formulate a mathematical "law of surfing," which says Web traffic follows predictable patterns.

But though the Internet Archive is helping tackle the challenges posed by the Web's ephemeral nature, its information is still being stored on hard disks that will soon be defunct. Within the past year, however, promising proposals have surfaced that could solve the obsolescence issue.

The most noteworthy is a records system created by the San Diego Supercomputer Center that converts digital documents into XML (Extensible Markup Language), a computer language that is gaining dominance on the Web. By using textual "tags" that tell what's what -- like where a title is, or where a graphic is -- XML frees documents from the software they're created in, allowing future computer systems to read them.

Unlike HTML (Hypertext Markup language), XML separates a document's content from its appearance, allowing it to be displayed as it originally looked when ran with the appropriate XML style program.

In less than two days, the Supercomputer Center converted one million White House e-mail messages into this universal format. The center also converted other federal documents like maps and Census data, proving their system could permanently preserve diverse and large volumes of electronic records that would be accessible decades hence, regardless of future technology.

In March, Carlin announced NARA and the National Science Foundation will study using this "major technological breakthrough" to build an electronic records archive. Carlin estimated it will take $130 million and five years to build such a facility.

Although the final solution to digital storage remains to be hammered out, one thing is certain: it will only be found if computer geeks, politicians and librarians unite, said Brian Cooper, a Stanford University computer science doctoral student also researching the problem. Cooper's proposed solution involves saving records on a computer that is linked via the Internet to other institutions, who then back up each others' archives in case corruption occurs.

"What we (computer scientists) do now has the potential to really revolutionize how people use computers," Cooper said. "Today they're something you just use but can't really rely on, but what we're working on is looking at making them into sort of a collective memory of the human race."
The scramble to save Web history

Obsolete equipment poses challenges for Internet historians

By Jeanene Harlick
STAFF WRITER

Old Web pages don't die, they just fade away.

In fact, web pages are blinking away so fast that the very history of these pioneer Internet years is threatened with extinction.

That's why a small group of scientists and students is racing to find ways to archive web pages, just as paper documents like the original Declaration of Independence have been preserved.

The difference, said archivist Jean Dekem, is that paper persists and, so far, objects in the digital world don't.

"They don't give any warning they're going to corrupt. They simply very quietly pass out of usefulness," said Dekem, who is with the Stanford Linear Accelerator Center.

A few years ago she found herself trying to electronically chase down the nation's first Web page, which was born in 1991 in a SLAC physicist's office.

No record of the page existed. Plowing through old backup tapes, the center's Web master found it, a page of plain text that made virtual history.

If the pair had waited one year, the page would have been lost forever, trapped in a file created by obsolete software no current hard drive can read.

It's too late, however, for the first e-mail ever sent. Too late, as well, for countless original World Wide Web pages and other documents stored electronically.

Confronted with voluminous electronic records, storage media that is obsolete within a few years and employees who delete documents before anyone realizes they're important, some fear future historians will look back on this time as a "digital dark ages."

"The chronicle of our entire period (is) in jeopardy," computer scientist Jeff Rothenberg wrote in a 1995 article that sent a shock wave through the digital community. "It is only slightly facetious to say that digital information lasts forever -- or five years, whichever comes faster."

Information overload

A look at the National Archives and Records Administration shows just one example of the depth of the problem.

Over the past quarter-century, the National Archives has been able to process just 100,000 e-records. But the Treasury Department alone is generating nearly one million e-mail messages a year that should be permanently preserved, U.S. Archivist John Carlin estimates.

In 1976, when the National Archives and Records Administration went looking for the 1960 U.S. Census data, officials realized it was stored on UNIVAC tapes whose tape drives -- the first main frame computers -- no longer existed. It took three years and some engineering back flips to eventually recover the data.

In addition, the National Archives has no digital archives or guidelines on how federal agencies should store electronic records the agencies are mandated to preserve.

"New technological ways to create records have not been matched by technological advancements to manage them," Carlin
Unlike HTML (Hypertext Markup language), XML separates a document's content from its appearance, allowing it to be displayed as it originally looked when ran with the appropriate XML style program.

In less than two days, the Supercomputer Center converted one million White House e-mail messages into this universal format. The center also converted other federal documents like maps and Census data, proving their system could permanently preserve diverse and large volumes of electronic records that would be accessible decades hence, regardless of future technology.

In March, Carlin announced NARA and the National Science Foundation will study using this "major technological breakthrough" to build an electronic records archive. Carlin estimated it will take $130 million and five years to build such a facility.

Although the final solution to digital storage remains to be hammered out, one thing is certain: it will only be found if computer geeks, politicians and librarians unite, said Brian Cooper, a Stanford University computer science doctoral student also researching the problem. Cooper's proposed solution involves saving records on a computer that is linked via the Internet to other institutions, who then back up each others' archives in case corruption occurs.

"What we (computer scientists) do now has the potential to really revolutionize how people use computers," Cooper said. "Today they're something you just use but can't really rely on, but what we're working on is looking at making them into sort of a collective memory of the human race."

© 2000 by MediaNews Group, Inc. and ANG Newspapers
Playmate visit U.S. OKs Enviro's

WASHINGTON — In the early days of space exploration, the idea of a mission to Mars was considered futuristic and science fiction. However, with advancements in technology and manned space flight, NASA is now planning to send a crew to Mars in the 2030s. This ambitious project poses significant challenges, including the development of suitable equipment and the psychological impact on the astronauts. The director of the Mars exploration program has expressed confidence in the program's success, stating that the agency is well-prepared and ready to take on this monumental task. By James Hyneman

Sent funds to take kids Employee allegedly accused of theft

The scramble to save Web history

The Internet and its impact on modern life is a topic of great interest. As the Web turns 21 in 1991, its birthplace at the Stanford Linear Accelerator Center in Palo Alto, California, has become a symbol of innovation and progress. The world of information is constantly evolving, and the Internet has played a crucial role in shaping our society. From email to e-commerce, the Web has transformed the way we communicate and conduct business. By Mark A. Mcに入った

I am a long
FIRST WEB PAGE Created for research
New weapon against heartburn goes down easy

Police chief of Colma fired

The scramble to save Web history

Netting relics

Supreme Court may kill ‘blanket prime’
History: Nonprofit Internet Archive started recording Web in 1996

First Web page created for research

Jeanenne Markles

When Vice President Al Gore talked up the idea of putting all government websites on the Internet, then-National Archives and Records Administration historian Jeanenne Markles saw a chance to preserve Web history.

"When I heard the idea, I thought it was a great idea," Markles said. "I thought it would be really interesting to see how Web pages evolve over time."

So in 1996, Markles and her team began gathering websites, not just government sites but any site that had something to do with computers or the Internet.

"We tried to include everything," Markles said. "It was a bit of a challenge to get permission to copy the sites." But Markles said they were able to negotiate with most of the sites they wanted to include.

In 1996, the Internet Archive started recording the Web, capturing the pages as they were created. The archive now contains more than 400 million pages from over 200 million websites.

"We want to make sure that the Web is preserved," Markles said. "So we capture the sites and make them available to the public for future generations."
HEADLINE NEWS

The scramble to save Web history

Old Web pages don’t die, they just fade away. In fact, web pages are blinking away so fast that the very history of these pioneer Internet years is threatened with extinction. That’s why a small group of scientists and students is racing to find ways to archive web pages, just as paper documents like the original Declaration of Independence have been preserved....More

BUSINESS

New Andronico’s project touts homestyle meals

EMERYVILLE -- A BANNER OUTSIDE the renovated warehouse on Hollis Street says "Andronico’s Top Secret Project," but tables and chairs sitting on the patio provide clues that this will be more than a small grocery store. Andronico’s officials believe that its so-called Central Market, scheduled to...More

SPORTS

Sharks set to battle Game 7-savvy Blues

ST. LOUIS -- Tony Granato has played in 713 NHL regular season games. He has played in 69 Stanley Cup playoff games. He has played in one-seventh game of a playoff series. It was 1993, and the San Jose Sharks’ veteran winger was a member of the Los Angeles Kings. The conference finals against...More

need a JOB?

MORE HEADLINE NEWS
Haven for needy families

MENLO PARK -- On the former site of a "no-tell" motel, a new building is taking shape that helps homeless families plant roots in permanent homes. The new Haven House is expected to open by the end of June, to provide transitional housing for 23 families at a time. The $4.....More

© 2000 by MediaNews Group, Inc. and ANG Newspapers
Ten years ago, Intel's 25-MHz 486 microprocessor was state of the art. The Internet was just a tool for universities, scientists, and the military. Windows 3.0 was still in development. Steve Jobs was in exile from Apple.

We've come a long way, baby.

It's time to reminisce and take stock of the '90s, a decade that has seen the computing industry become a driving force in the world economy. This is the first entry in a four-part CNET series saluting the most important people, products, trends, and success stories in computing over the last decade.

Part 1 focuses on visionaries, the people whose ideas have had a profound impact on computing in the 1990s--and whose influence will be felt well into the next century. Some of them are business leaders; others are engineers and programmers. And they've all changed the world.

How did we pick them? We made a list of everyone who had a great idea or engineered a huge success at some point in the last ten years. Then we argued a lot until we whittled the list down to ten. These are CNET's picks for the ten most influential visionaries of this decade.

Visionary No. 10: Clark Kent or Superman?


Valerie Potter is senior editor and Matt Rosoff is senior writer for CNET Special Reports. Rex Baldazo, a content presentation engineer for CNET Technology, also contributed to this story. Got a comment? Let us know!

Want to share your comments with other readers? Go tell it in CNET Message Board!
Marc Andreessen
Coinventor of the integrated text-and-graphics Web browser and cofounder of Netscape

Contrary to popular opinion, Marc Andreessen did not invent the Web browser. Nor did he create the first browser with a graphical user interface. Andreessen's browser wasn't even the first to use pictures. Nevertheless, the Web wouldn't be where it is today without him.

So what exactly did Andreessen do? He made the Web accessible to the masses.

Who paved the way?
The most important of Andreessen's predecessors was Tim Berners-Lee, who invented the Web and its first browser, WorldWideWeb, for NeXT. Several others developed important early browsers:

- Nicola Pellow wrote a text-based browser called line mode or www that would run on almost any computer.
- Wei Pei wrote ViolaWWW for Unix.
- Kim Nyberg, Tenuu Rantanen, Kati Suominen, and Kari Sydanmaanlakka wrote Erwise for Unix.
- Tony Johnson wrote Mids for Unix.
- Tom Bruce wrote Cello for the PC.

And Netscape owes much of its success to Jim Clark, who cofounded the company, and Jim Barksdale, who led it to greatness.

Making browsers for fun...and profit
In early 1994, Andreessen joined up with Jim Clark, the founder of Silicon Graphics, to form the company that would become Netscape Communications. Andreessen brought most of his NCSA colleagues with him, and within months they released Mosaic Netscape (later renamed Netscape Navigator), a faster, slicker, more secure browser than the original Mosaic. Navigator was a runaway
success, quickly snapping up more than 70 percent of the browser market share. *Time* magazine even named it one of the ten best products of 1994, right up there with the Chrysler Neon, the Wonderbra, and Frutopia. (OK, so maybe that wasn’t such an honor.)

But the mythos surrounding the company was almost more important to the evolution of the Web than the browser itself. Along with Yahoo and other Web startups, Netscape embodied Web culture—young, hip, smart, irreverent—and the baby-faced, often barefooted Andreessen served as its poster child. The company generated so much buzz that its summer 1995 initial public offering stock, which had been valued at $28 a share, instead opened at $71—unbelievable for a company that had never turned a profit and gave most of its software away. Once again, Netscape was a leader: this time of the sky-high Internet IPO.

**Lasting impressions**

We all know what happened next. Microsoft came along and spoiled the party. After fending off the software giant for years with increasingly desperate measures, Netscape finally threw in the towel in November 1998 and agreed to be bought out by America Online. But not before it had made its mark on the Web forever by introducing advanced HTML tags (some more successful than others), developing the widely used JavaScript coding language, and incorporating third-party innovations like Java into the browsing experience (and forcing Microsoft to do the same).

Andreessen has continued to promote the Web agenda over the years, championing technical standards, the open source movement, and innovative new technologies. As he moves to his new position as chief technology officer of AOL, those who remember the Web's pioneer days hope he will take that spirit with him.

**Visionary No. 2: who's afraid of the big, bad wolf?**

---

*Advertisement*

**Software from Beyond.com**

Save on Turbo Tax and Quicken • Antivirus Software! • Visit the Millennium Center!