

Network Based Documents

Sandy Ressler

National Institute of Standards and Technology

October 1994

Abstract

Networks of computers are changing the traditional roles of author, manufacturer, publisher, and distributor. A document written for one purpose can have many lives. Electronic documents can be authored, maintained, distributed and read on-line. The line between documents and data bases is blurred. Network browsing tools give users new freedom to explore mass quantities of information. Often the task at hand is to locate information.

Network based tools come in many varieties. Searching engines, worms, robots and browsers are a few of the names associated with this growing class of information technology applications. In this chapter we will examine ways in which documents intended to reside on networks effect the roles one must play both as author and user of information.

1.0 Introductory Questions

The best place to start examining the issues is to ask some questions. What is a document in the context of electronic authoring and distribution? How can I find the information I want in the vast swamp of information out on the net? As an author what can I do to make the information in my documents more usable and useful?

The entire concept of a document is becoming increasingly obsolete. The document as database on a network cries for a new term (in the best Ted Nelson style) the *docubasenet*. The docubasenet has characteristics common to documents and databases and yet is something different. Most documents are meant to be printed, on paper. The look of words the typography and layout is important. The design of a document and the rela-

relationship of the words to the illustrations by their arrangement serve to enhance the user interface of the document. A document pleasing to the eye will convey its content better than a document that is visually difficult to read.

Databases, information containers, are made accessible via a myriad of techniques; forms, full-text queries and batch processes. The information in a database is useless without a mechanism to get to the data in useful ways.

The docubasenet is information organized in a structured way, like a database, presented to the user as a document which can be queried. Distribution of the docubasenet occurs via networks. The distributed nature of information available via wide area networks opens up a wide range of options, possibilities and problems.

1.1 Writing

Writing for a networked distributed environment changes the process of creating information in some subtle and not so subtle ways. First the structure of the information becomes much more important. Information that is designed to be accessed via queries must be organized in a structured manner. The better and more organized the structure the better the chances that queries will return meaningful information. One side effect of this is the interest in the Standard Generalized Markup Language [SGML]. SGML enables one to create highly structured documents. Writing is transformed into authoring. We see the creation and marketing of authoring systems which enable the information creator to format and structure the information in new ways suitable for distribution on networks and viewing with new multimedia tools.

1.2 Distribution

The printing and distribution process is greatly modified in a networked environment. One can simply copy files, or as we will see later, browse documents across a network. The ease with which one can move information across networks cuts both ways however. The legal and cost recovery issues associated with author and publisher rights are still a major problem with no uniform legal solution. The issues in this case are not primarily technical; they are legal.

The more open and widespread a network becomes, the more useful and problematic it becomes. Let's now examine the "Mother of all Networks," the Internet.

2.0 Internet background

2.1 Introduction to the internet

The Internet, which I refer to simply as the net, grew up from an interconnected set of research institutions originally funded by the United States Department of Defense and known as the ARPANET. As the importance of this connectivity grew the non-defense oriented research sector of the United States government started to support the net, principally the National Science Foundation (NSF). Currently the net is in a state of extreme change. The commercial world has jumped on the Internet bandwagon and commercial Internet service providers offer a wide variety of Internet connectivity options.

2.2 Getting Connected

There are an increasing variety of choices to get connected to the Internet. Large on-line service providers, such as America On-line now offer “full” Internet connectivity. As a consumer you must watch carefully for what the notion of “full connectivity” entails. Email, telnet and ftp access are parts of the equation. Information applications such as Gopher and Mosaic are an increasingly important part. One of the most intriguing products is O’Reilly’s Internet In A Box [IIBOX].

According to some O’Reilly marketing literature:

Internet In A Box is the first shrink-wrapped package to provide a total solution for PC users wanting to get on the Internet. Internet In A Box provides instant connectivity, a multimedia Windows interface, a full suite of applications, and the first interactive guide to the Internet.

- *A subscription to the Global Network Navigator (GNN), the first interactive guide to the vast information resources of the Internet (viewed using Mosaic).*
- *Software: The Air Series of Internet-access tools, including Mosaic, electronic mail, Usenet news reader, drag-and-drop file transfer, and telnet.*
- *Two books that clearly describe how to use these resources: “A Guide to Getting Started”, and a special edition of Ed Krol’s bestselling book, “The Whole Internet User’s Guide & Catalog.”*

As the bandwidth demands increase, simple, even high speed, modems become insufficient. Multimegabyte images and hundred megabyte videos demand higher capacity networks. The solution just might be ISDN, the Integrated Services Digital Network [ISDN]. ISDN seems to be the network solution of choice by the most ubiquitous network providers, the telephone companies. There are even companies starting to offer “ISDN on-ramps” to the Internet. PSI’s InterRamp [PSI] starts at about \$30/month for 64K baud access. While configurations vary a typical ISDN configuration is usually 128K baud for a ball-park cost of \$100-150/month; not a trivial cost but much less than the costs of dedicated T1 lines.

2.3 Commercializing the Net

In addition to the increasingly commercial network connectivity services the information content providers are starting to look at the net as a distribution medium. Publishers are generally in business to make a profit. One of the most significant problems facing network publishers is the lack of a secure reliable method to bill customers. Sending charge card numbers across networks is not a comforting prospect to most people. The answer to security is reliable trustworthy encryption techniques. A number of different methodologies are available and in development to solve the security issue, at least from a technical point of view.

The most highly touted security technique is public key encryption [KEY]. Public key encryption involves the use of two keys. One is public and known to everyone; the other is private and known only to the message sender. Various techniques using the keys

enable digital signatures, document authentication, privacy and other security mechanisms.

3.0 Resource Discovery Tools

As the Internet becomes larger and more interconnected. Finding information is becoming one of the most challenging issues. Often one is struck by thoughts such as “I know I saw something about the new netWizard product...but where was it?” An entire discipline has been created called resource discovery [SCHWARTZ]. Resource discovery is the formal study of net surfing:-).

3.1 Archie

Some very clever people at McGill University in Canada have created one of the earliest and most accessible means of locating information on the Internet. They created an electronic mail response program called “archie.” Archie maintains a database of archive sites and the names of their contents. The command used for searching functions follows:

```
prog <reg expr1> [<reg exp2>...]
```

`prog` is a keyword which means to find or search. A search of the “archie” database is performed with each `<reg exp>` (a regular expression) in turn, and any matches found are returned to the requestor. Note that multiple regular expressions may be placed on one line, in which case the results will be mailed back to you in one message. If you have multiple “prog” lines, then multiple messages will be returned, one for each line.

A friendlier user interface exists via the program `xarchie`. An X window system program, `xarchie` lets you interactively select the database sources and pose queries. `xarchie` returns the results of the query in a list with the most relevant items on the top of the list.

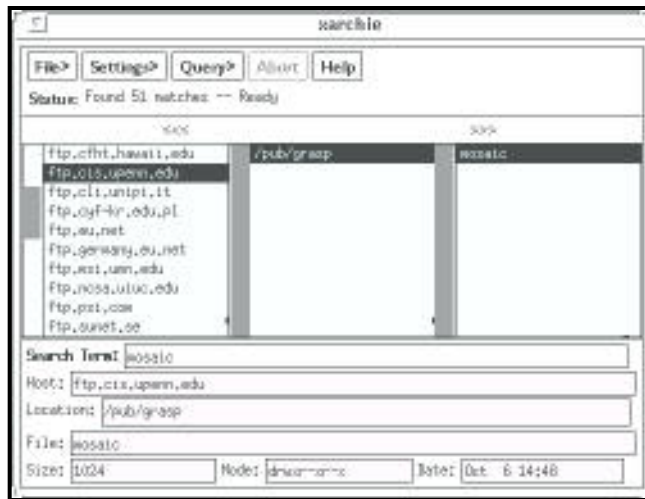


FIGURE 1. XARCHIE - the X Window System User Interface toarchie

Users ofarchie simply send email requests containing the command `prog` using the syntax specified above to search for particular programs or documents of interest. Archie email sends a response, telling you where to go to find the item you requested.

Anarchie site runs a program that maintains the database by using anonymous ftp. Thearchie program that build and maintains the database checks the contents of several hundred archive sites over approximately 1 to 2 months. Several sites now runarchie servers. Some sites allow interactive queries of thearchie database via TELNET, eliminating the delay inherent in email.

3.2 WAIS

The Wide Area Information Servers (WAIS) system points to one of the most significant advances of the networking community. WAIS is an effort to make possible network-wide document retrieval. Keep in mind that information queries travel across a network such as the Internet. The project is a joint effort of Thinking Machines Corporation, Dow Jones News/Retrieval, and Apple Computer. This project is clearly pointing the way to the future of information access.

The following overview (written in April 1991) is from the WAIS project leader Brewster Kahle:

The Wide Area Information Servers system is a set of products supplied by different vendors to help end-users find and retrieve information over networks. Thinking Machines, Apple Computer, and Dow Jones initially implemented such a system for use by business executives. These products are becoming more widely available from various companies.

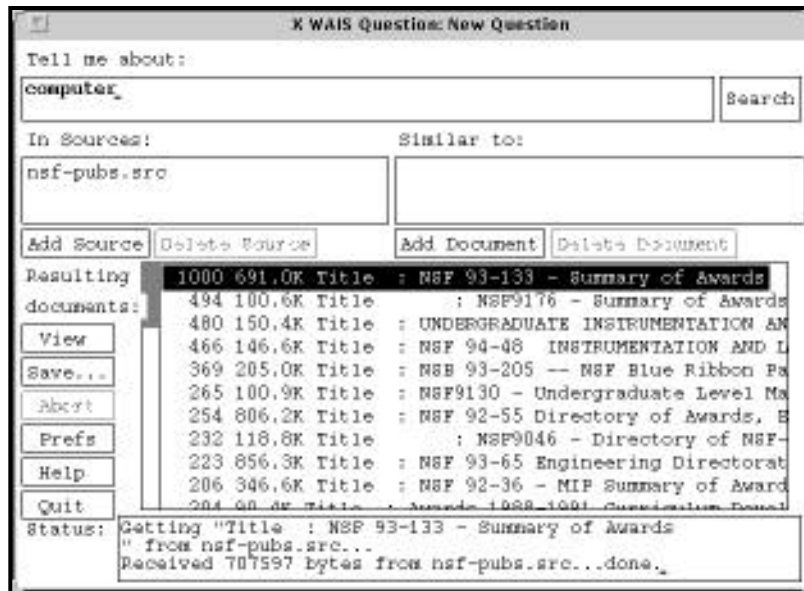


FIGURE 2.

XWAIS - the X Window System User Interface to WAIS, results of a query

What does WAIS do? Users on different platforms can access personal, company, and published information from one interface. The information can be anything: text, pictures, voice, or formatted documents. Since a single computer-to-computer protocol is used, information can be stored anywhere on different types of machines. Anyone can use this system since it uses natural language questions to find relevant documents. Relevant documents can be fed back to a server to refine the search. This avoids complicated query languages and vendor specific systems. Successful searches can be automatically run to alert the user when new information becomes available.

How does WAIS work? The servers take a user's question and do their best to find relevant documents. The servers, at this point, do not "understand" the user's English language question, rather they try to find documents that contain those words and phrases and rank them based on heuristics. The user interfaces (clients) talk to the servers using an extension to a standard protocol Z39.50. Using a public standard allows vendors to compete with each other, while bypassing the usual proprietary protocol period that slows development. Thinking Machines is giving away an implementation of this standard to help vendors develop clients and servers.

What WAIS servers exist? Even though the system is very new, there are already several servers:

* Dow Jones is putting a server on their own DowVision network. This server contains the Wall Street Journal, Barons, and 450 magazines. This is a for-pay server.

** Thinking Machines operates a Connection Machine on the internet for free use. The databases it supports are some patents, a collection of molecular biology abstracts, a cookbook, and the CIA World Factbook.*

** MIT supports a poetry server with a great deal of classical and modern poetry. Cosmic is serving descriptions of government software packages. The Library of Congress has plans to make their catalog available on the protocol.*

** Weather maps and forecasts are made available by Thinking Machines as a repackaging of existing information.*

** The "directory of servers" facility is operated by Thinking Machines so that new servers can be easily registered as either for-pay or for-free servers and users can find out about these services.*

How can I find out more about WAIS?

Contact Brewster Kahle for more information on the WAIS project, the Connection Machine WAIS system, or the free Mac, Unix Server, and X Window System interfaces. There is a mailing list that has weekly postings on progress and new releases; to subscribe send an email note to wais-discussion-request@think.com.

*Brewster Kahle
Project Leader Wide Area Information Servers
Brewster@Think.com*

It is important to note that the communications to the WAIS servers are accomplished using the ANSI standard protocol for database retrieval applications, Z39.50. The decision to use a public standard is what makes this communications method truly open.

3.3 Gopher

Gopher is a widely used method to browse through documents on the Internet. Originated at the University of Minnesota, gopher usage has exploded.

One way of locating information on the Internet is to use gopher servers. Gopher servers maintain collections of documents with the additional ability of full-text searching. The gopher protocol and concept are due to the effort of the people at the Microcomputer and Workstation Networks Research Center at the University of Minnesota. One of the appealing aspects of gopher is the simplicity with which it presents itself to the user. You, the user, are presented a "file system" just like any hierarchically organized file system, except that this file system covers all information known to the particular gopher server. It's a simple, elegant, and powerful approach.

In the case of an interaction with gopher servers, however, these directories may exist anywhere on the Internet. Furthermore, the user doesn't really care where the information is, as long as it's accessible in a timely manner.

Client systems through which a user “speaks” to the gopher server can have a number of user interfaces. One of the clients on UNIX systems is based on `curses` and will function on any terminal.



FIGURE 3.

Typical Gopher menu

According to Mark McCahill member of the gopher development team:

The Internet Gopher is a distributed document delivery service. It allows a neophyte user to access various types of data residing on multiple hosts in a seamless fashion. This is accomplished by presenting the user a hierarchical arrangement of documents, a menu, and by using a client-server communications model. In addition to browsing through hierarchies of documents, gopher users can submit queries to gopher search servers. The search servers typically have full-text indexes for a set of gopher documents; the response to a query is a list of documents that matched the search criteria.

Internet Gopher servers accept simple queries (sent over a TCP connection), and respond by sending the client a document or a list of documents. Since this is a distributed protocol there can be many servers... but the client software hides this fact from the user. We currently use this technology at the University of Minnesota to help support microcomputer users... a couple of gopher servers have 6000-7000 computer Q&A items that users can search for answers to their questions. In addition, there are also gopher servers with recipes and other fun stuff.

3.3.1 VERONICA

VERONICA, the Very Easy Rodent-Oriented Net-wide Index to Computerized Archives, is a utility which indexes gopher items. From the Veronica FAQ (Frequently Asked Questions) comes the following:

Q1: What is veronica?

A1: Veronica is a service that maintains an index of titles of gopher items, and provides keyword searches of those titles. A veronica search originates with a user's request for a search, submitted via a gopher client. The result of a veronica search is a set of gopher-type data items, which is returned to the gopher client in the form of a gopher menu. The user can access any of the resultant data items by selecting from the returned menu. A veronica search typically searches the menus of hundreds of gopher servers, perhaps all the gopher servers that are to the Internet.

At present, there are no "veronica clients" per se; veronica is accessed through normal gopher clients. Veronica is tightly integrated with the gopher protocol.

The veronica service comprises two functions:

- 1). Harvesting menu data from gopher servers, and preparing it for se;*
- 2). Offering searches of that database to gopher clients.*

These two functions are not necessarily provided by the same host computer. Most users and administrators of veronica search servers will not need to be concerned with the first phase of the process. Operators of veronica query-engines can obtain a prepared dataset for use with the query server.

Veronica evolved as a solution to the problem of resource discovery in the rapidly-expanding gopher meta-burrow. At the University of Nevada, there was an outcry for an easy way to find gopher-based information without doing a menu-by-menu, site-by-site search.¹

3.3.2 JUGHEAD

JUGHEAD, Jonzy's Universal Gopher Hierarchy Excavation And Display, is a utility to get menu information from gopher servers. It allows you to search for specific gopher spaces. The following information comes from the About.jughead file included with the jughead distribution²:

jughead head can act as a search engine on a prebuilt table that allows searching through menus, or can create a linear view of menu space. When running jughead you can specify what part of gopherspace you want search tables built or a linear view thereof.

When running as a search engine, jughead listens out a port for a connection and a search string. The search string can contain the boolean operations "AND", "OR", and "NOT" between words. If no operator is specified between words, it is an implied "AND" operation. For example either:

-
1. Veronica FAQ maintained by Steven Foster and Fred Barrie, version 1994/02/08.
 2. Available via anonymous ftp from ftp.cc.utah.edu in /pub/gopher/GopherTools

"University of utah" or "University AND of AND Utah"

will yield all gopher entries with "university" and "of" and "utah" in the title. The case of the letters is irrelevant. Now suppose you enter the string:

"university of utah NOT gopher"

This will return the same information as the first search except for those entries containing the word "gopher".

3.4 Worms and Knowbots

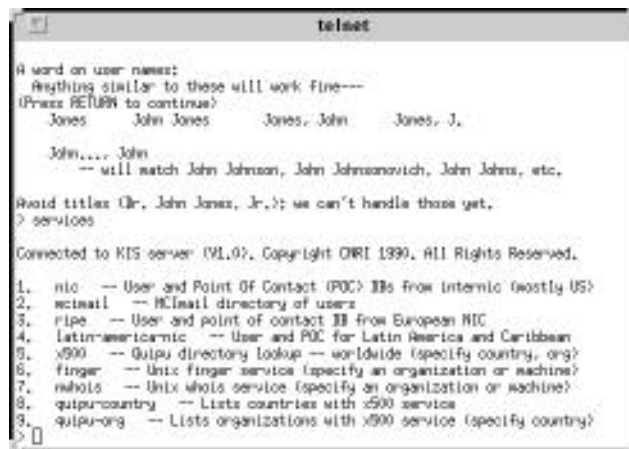
Web worms and Knowbots are automated tools which crawl around the web looking for information, reporting their findings. A worm is a program that moves from one site to another. The generic term "worm" has nothing to do with the WWW and simply refers to a program that seeks to replicate itself on multiple hosts. Worms are not necessarily good. The "Internet Worm" of 1988 caused a massive breakdown of thousands of systems on the Internet.

A knowbot is a program or agent that, like worms, travels from site to site, although with a more artificial intelligence flavor usually following knowledge-based rules. Another term for a knowbot might be an autonomous agent. Clear distinction between these terms are currently not meaningful. In the context of this section, finding information, let's look at one particular knowbot and one worm.

3.4.1 Knowbot

Finding email addresses for people is one of the more frustrating tasks one might attempt on the Internet. There are a number of tools to aid this process. Actually there are too many tools. Sorting through a collection of tools is the CNRI Knowbot. It provides a homogeneous front end to MCI mail, x.500, finger, nwhois and other people finding tools.¹

1. The CNRI Knowbot is located at the URL <telnet://info.cnri.reston.va.us:185>



```
telnet
A word on user names:
Anything similar to these will work fine---
(Press RETURN to continue)
  Jones   John Jones   Jones, John   Jones, J.

  John... John
  -- will match John Johnson, John Johnsonovich, John Johns, etc.

Avoid titles (Mr. John Jones, Jr.): we can't handle those yet.
> services

Connected to KIS server (VL0). Copyright CNRI 1990. All Rights Reserved.

1. nro -- User and Point of Contact (POC) IDs from Internic (mostly US)
2. netmail -- Netmail directory of users
3. ripe -- User and point of contact ID from European NIC
4. latin-amer-icnic -- User and POC for Latin America and Caribbean
5. x900 -- Quipu directory lookup -- worldwide (specify country, org?)
6. finger -- Unix: finger service (specify an organization or machine?)
7. whois -- Unix: whois service (specify an organization or machine?)
8. quipecountry -- Lists countries with x900 service
9. quipe-org -- Lists organizations with x900 service (specify country?)
>
```

FIGURE 4. Telnet interaction with CNRI Knowbot

3.4.2 The WebCrawler

Following is a description of the WebCrawler one particular Web worm available at <http://www.biotech.washington.edu/WebCrawler/WebCrawler.html>.

The WebCrawler is a web robot, and is the first product of an experiment in information discovery on the Web. I wrote it because I could never find information when I wanted it, and because I don't have time to follow endless links.

The WebCrawler has three different functions:

It builds indices for documents it finds on the Web. The broad, content-based index is available for searching. It acts as an agent, searching for documents of particular interest to the user. In doing so, it draws upon the knowledge accumulated in its index, and some simple strategies to bias the search toward interesting material. In this sense, it is a lot like the Fish search, although it operates network-wide. It is a testbed for experimenting with Web search strategies. It's easy to plug in a new search strategy, or ask queries from afar, using a special protocol.

In addition, the WebCrawler can answer some fun queries. Because it models the world using a flexible, OO approach, the actual graph structure of the Web is available for queries. This allows you, for instance, to find out which sites reference a particular page. It also lets me construct the Web Top 25 List, the list of the most frequently reference documents that the WebCrawler as found.

How it Works

The WebCrawler works by starting with a known set of documents (even if it is just one), identifying new places to explore by looking at the outbound links from that document, and then visiting those links.

It is composed of three essential pieces:

The search engine directs the search. In a breadth-first search, it is responsible for identifying new places to visit by looking at the oldest unvisited links from documents in the database. In the directed, find-me-what-I-want strategy, the search engine directs the search by finding the most relevant places to visit next. The database contains a list of all documents, both visited and unvisited, and an index on the content of visited documents. Each document points to a particular host, and, if visited, contains a list of pointers to other documents (links). "Agents" retrieve documents. They use CERN's WWW library to retrieve a specific URL, then returning that document to the database for indexing and storage. The WebCrawler typically runs with 5-10 agents at once.

Being a Good Citizen

The WebCrawler tries hard to be a good citizen. Its main approach involves the order in which it searches the Web. Some web robots have been known to operate in a depth-first fashion, retrieving file after file from a single site. This kind of traversal is bad. The WebCrawler searches the Web in a breadth-first fashion. When building its index of the Web, the WebCrawler will access a site at most a few times a day.

When the WebCrawler is searching for something more specific, its search may narrow to a relevant set of documents at a particular site. When this happens, the WebCrawler limits its search speed to one document per minute, and sets a ceiling on the number of documents that can be retrieved from the host before query results are reported to the user. The WebCrawler also adopts several of the techniques mentioned in the Guidelines for Robot Writers.¹

Implementation Status

The WebCrawler is written in C and Objective-C for NEXTSTEP. It uses the WWW library from CERN, with several changes to make automation easier. Whenever I feel comfortable about unleashing the WebCrawler, I'll make the source code available!

bp@cs.washington.edu

Brian Pinkerton

1. Guidelines for Robot Writers is available at <http://web.nexor.co.uk/mak/doc/robots/guidelines.html>.

4.0 Mosaic and the World Wide Web

4.1 Introduction to Mosaic

As others have stated, Mosaic is the first Internet “killer” Internet application. Mosaic was introduced to the net in the same way as most other university research projects. It is available for free with source code for not-for-profit use. It, like many other applications is a product of the Internet community, specifically the National Center for Supercomputing Applications (NCSA). The developers of Mosaic did not seek to invent everything. They built upon a number of existing standards and systems. Prime among these is the World Wide Web (WWW), developed at CERN. In fact, most of the technological “break-throughs” are the result of WWW. The fuss and hooplah surrounding Mosaic is due to the unified and reasonably pleasant interface it presents to the user.

Mosaic and it’s commercial clones, such as Netscape from Netscape Communications, offer end users a view of a compound document. Many items in the document contain links to other documents. These hypertext links allow the user to easily browse an entire collection of related documents. The documents are distributed and accessed throughout the Internet via the protocols supported by the World Wide Web. The net effect (pun intended) is to be able read compound documents with images and sounds with information dispersed over the network. Mosaic has become a usable front end to the Internet.



FIGURE 5. NCSA Mosaic and Mosaic Communication’s Netscape

According to the WWW FAQ (Frequently Asked Questions) maintained by Thomas Boutell:

What are WWW, hypertext and hypermedia?

WWW stands for "World Wide Web". The WWW project, started by CERN (the European Laboratory for Particle Physics), seeks to build a distributed hypermedia system.

The advantage of hypertext is that in a hypertext document, if you want more information about a particular subject mentioned, you can usually "just click on it" to read further detail. In fact, documents can be and often are linked to other documents by completely different authors -- much like footnoting, but you can get the referenced document instantly!

To access the web, you run a browser program. The browser reads documents, and can fetch documents from other sources. Information providers set up hypermedia servers which browsers can get documents from.

The browsers can, in addition, access files by FTP, NNTP (the Internet news protocol), gopher and an ever-increasing range of other methods. On top of these, if the server has search capabilities, the browsers will permit searches of documents and databases.

The documents that the browsers display are hypertext documents. Hypertext is text with pointers to other text. The browsers let you deal with the pointers in a transparent way -- select the pointer, and you are presented with the text that is pointed to.

Hypermedia is a superset of hypertext -- it is any medium with pointers to other media. This means that browsers might not display a text file, but might display images or sound or animations.

Mosaic is a WWW browser with hypermedia capabilities. Mosaic documents are a form of compound documents. The compound document a user manipulates is "authored" using the HyperText Markup Language (HTML) which is a specific Document Type Definition (DTD) of the Standard Generalized Markup Language (SGML). In short the WWW designers wisely chose not to invent yet another language technology and instead chose an existing standardized language. The developers of Mosaic used the rich foundation of WWW as a starting point. These types of collaborations are what makes an open Internet such a valuable resource.

4.2 Authoring

Using the WWW browser is easy, authoring documents for WWW publishing is relatively easy but sometimes a little tricky. WWW Documents must be in the HTML format. HTML is a specific application of SGML. Given that HTML is SGML one can purchase commercial off-the-shelf SGML products which can help in the writing and analysis of WWW documents.

In fact, the traditional SGML vendor community is jumping on the WWW as a new found market. Vendors are releasing SGML products specifically to support and aid the

writing of HTML documents. The decision by WWW developers to use SGML as the foundation for WWW documents is one of the Web's great strengths.

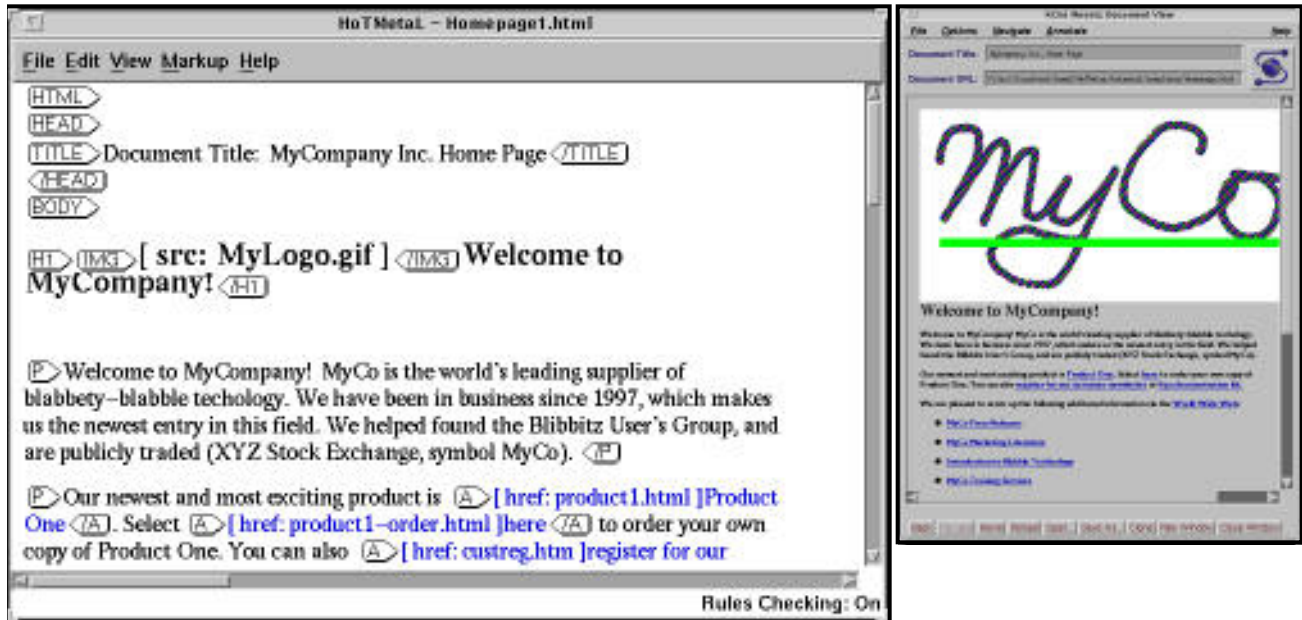


FIGURE 6. SoftQuad's HoTMetaL HTML editor and corresponding Mosaic view of document

HoTMetaL runs on PCs and Sun workstations. In addition there exists a number of free-ware packages for PCs and Macintosh. SimpleHTML is a HyperCard based editor for the Mac.

On UNIX workstations there is also tkWWW, an authoring tool based on the tk toolkit for X windows.

One of the most common methods is to simply use a text editor to write HTML directly. This "Iron Man HTML" technique can be aided by some editors, for example emacs, which provides an HTML mode to take some of the tedium away. HTML documents are readable and editable, they just require a little learning.

```

<!DOCTYPE HTML SYSTEM "html.dtd">
<HTML><HEAD><TITLE>Open Virtual Reality Testbed</
TITLE></HEAD></HTML>
<H1>Open Virtual Reality Testbed Home Page</H1>
<A HREF="OVRTbrochure.pdf"><IMG
SRC="ovrtlogo2G.gif"><IMG SRC="pdfLogo.gif"
ALIGN="left"
> PDF file of OVRT brochure</A>
<HR>
<P>
    
```

FIGURE 7. An HTML fragment

An interesting authoring issue concerns the conformance of HTML files with the HTML DTD. Some authoring packages help force conformance and others let you get

away with sloppy HTML. Most browsers do not bother to check for compliant HTML and simply do the best they can with the display. If you are concerned with interoperability then it will pay in the long run to be in conformance.

An alternative to native HTML authoring is to write documents using your favorite text-editor, word process or publishing system and then convert to HTML. A large number of conversion programs exist for just this purpose. For example the rtf2html program on UNIX platforms will convert documents in the RTF format into HTML. Similarly there are converters for latex and FrameMaker. All of these conversions mechanisms depend on a “properly” written original document. For example documents written in WordPerfect must use the style feature in order to successfully convert. Often conversion mechanism require some hand editing to touch up conversion errors.

Following this conversion type of authoring Microsoft is introducing an add-on module for MS Word which will output HTML and Interleaf is producing Cyberleaf a higher-end HTML authoring tool which can read in WordPerfect, Word RTF, FrameMaker MIF, Interleaf and ASCII formats.¹

4.3 Maintenance

As the web of HTML documents grows, maintenance of the links in the documents becomes an issue. It is frustrating to select a link only to have the browser return an error that the document doesn't exist.

One particularly useful and freely available tool is the html-analyzer². Created by James Pitkow at the University of Colorado, it validates the html links. According to documentation accompanying the software the analyzer performs as follows:

Specifically, all.html files are processed to 1) validate the existence of anchors 2) check that all tags used have links associated with them throughout the database and 3) check for a one-to-one relation between links and tags. It is important to note that the validation process checks http & relative addressed links only.

4.4 System Dependencies

WWW browsers are applications which run on the user's client machine. The client operating system and particular configuration of the client software and networking all play a role in the operation and behavior of the application. The availability of ancillary applications and properly configured system-wide protocols contribute to the portability, or lack thereof, of the final document.

One important issue associated with wide distribution of HTML documents derives from the loosely coupled nature of the WWW browser with various applications commonly known as helper applications. On all platforms, Mosaic launches another helper application when the user encounter an image file. The particular application launched

1. Cyberleaf information is from the Nov. 21, 1994 Seybold Report on Desktop Publishing.

2. The html analyzer is available via anonymous ftp from ftp.uiuc.edu in the directory Mosaic/Contrib.

is dependent on the particular MIME¹ type of the data, and is often dependent on the extension used for the file name. If, for example, the HTML document points to a JPEG² formatted image, the client machine must have an application capable of displaying JPEG images and Mosaic must be configured to launch that application upon links to JPEG images. This same scenario occurs for sound and video files.

Documents created on one machine should not contain absolute path names. Authors should be careful to use only links with relative address names to other files on the same server. In this way the entire hierarchy of HTML files can be moved as a unit without concern for renaming file path names inside the documents. In addition the relative names often must only be names in directories down from the current location. This is a security feature.

As for security, WWW servers and Mosaic present a whole range of security problems.³ Some of the many security issues are authentication of requests and server, privacy of request and response. In particular one very useful capability is the ability to execute arbitrary programs through the server. The CGI (Common Gateway Interface) attempts to control how program execution occurs, but clearly security can be compromised.

The behavior and assumption used by one browser may be different from another resulting in documents that look different. One browser may automatically interpret a newline as a new paragraph and another may not.

5.0 Underlying Protocols/Standards

One of the great strengths of the WWW is the use of standards. The following alphabet soup of standards from the formal standards world and the Internet standards world are part of the technical foundations of the WWW.

5.1 URI

The URI⁴ or Universal Resource Identifier is a generic all-encompassing term used to identify all Uniform Resource (UR) specifications. Other UR specifications are by definition part of the set of URs described by the URI.

In addition there is ongoing work on the URC, a Uniform Resource Citation. A URC will be a set of attribute/value pairs describing an object. Some of the attributes include author, publisher, date, and copyright status.

1. MIME the Multipurpose Internet Mail Extensions specification is discussed in the section Underlying Protocols/Standards.

2. JPEG is the Joint Photographic Experts Group standard for image compression.

3. See <http://info.cern.ch/hypertext/WWW/Protocols/HTTP/Security.html> for pointers to lots of security information.

4. A good place to find out about UR specifications is <http://info.cern.ch/hypertext/WWW/Addressing/Addressing.html> by Tim Berner-Lee

5.2 URL

The Uniform Resource Locator (URL) specification is a way of naming and addressing objects on the net. Following is the Abstract from the Internet Draft specification of Uniform Resource Locators (URL):

Internet Draft - CERN - Uniform Resource Locators (URL)

A Unifying Syntax for the Expression of Names and addresses of Objects on the Network

Abstract

Many protocols and systems for document search and retrieval are currently in use, and many more protocols or refinements of existing protocols are to be expected in a field whose expansion is explosive.

These systems are aiming to achieve global search and readership of documents across differing computing platforms, and despite a plethora of protocols and data formats. As protocols evolve, gateways can allow global access to remain possible. As data formats evolve, format conversion programs can preserve global access. There is one area, however, in which it is impractical to make conversions, and that is in the names and addresses used to identify objects. This is because names and addresses of objects are passed on in so many ways, from the backs of envelopes to hypertext objects, and may have a long life.

A common feature of almost all the data models of past and proposed systems is something which can be mapped onto a concept of "object" and some kind of name, address, or identifier for that object. One can therefore define a set of name spaces in which these objects can be said to exist.

Practical systems need to access and mix objects which are part of different existing and proposed systems.

This paper discusses the requirements on a universal syntax which can be used to encapsulate a name in any registered name space. This will allow names in different spaces to be treated in a common way, even though names in different spaces have differing characteristics, as do the objects to which they refer.

The universal syntax to objects available using existing protocols, and may be extended with technology. It makes a recommendation for a generic syntax, and for specific forms for "Uniform Resource Locators" (URLs) of objects accessible using existing Internet protocols.

The syntax has been in widespread use by World-Wide Web software since 1990.

5.3 URN

The Uniform Resource Name (URN) associated with an item is meant to be a persistent name which would return a list of current URLs pointing to the item. The URN is assigned by an Internet naming authority, such as the IANA (Internet Assigned Number Authority). In an example taken from a recent Open Systems Today article:¹

the URL of a file

`ftp://ftp.uu.net/published/osys-today/urdrafts.tar.Z`

and the URN for the same file could be:

`URNg1`

`IANA:merit.edu:9283492.`

In the URN, the IANA indicates that the following organization (merit.edu) has the authority to assign URN IDs and that the following number is the ID assigned by that organization.

Searching clients such as WAIS which use URN would return a list of URLs, one for each site containing the document. Over time sites may delete or add the document, but the URN would always return the most recent list of URLs.

5.4 SGML

The Standard Generalized Markup Language (SGML) has grown in popularity and use recently corresponding to the growth of the WWW. HTML, a specific SGML application, has fueled this growth and awareness of SGML. SGML is a markup language used primarily to define and markup the structure of a document [SGML].

Logical document structures are the components of a document such as chapters, sections, heading and paragraphs, which comprise the entire document. Instances of these logical document items are the document itself. For example the name <CHAP> might refer to all chapters, a structural item of a document, and “*Chapter 3 - Web Feet*” is a particular chapter.

The entire structure of a document is defined in a Document Type Definition (DTD) which is a kind of metalanguage. DTDs define the structure of documents in a rigorous formal manner. Once a DTD is defined authors can write documents which conform to the DTD. In the case of the WWW, authors create documents which conform to the HTML DTD. Applications do not necessary enforce conformance which often causes confusion as one person’s document “worked just fine” and when viewed on another browser has problems. Indeed the HTML DTD did not, for a time, keep pace with HTML features.

1. “Working Group Starts to Put Together A Guide for A Standard Internet Landscape” by Jason Levitt, Open Systems Today Mar. 28, 1994.

5.5 Z39.50

A protocol for search and retrieval tasks usually associated with the library information retrieval community. A glossary entry¹ explaining Z39.50 states that the Z39.50 protocol is “Name of the national standard developed by the National Information Standards Organization (NISO) that defines an applications level protocol by which one computer can query another computer and transfer result records, using a canonical format. This protocol provides the framework for OPAC (On-line Public Access Catalog), users to search remote catalogs on the Internet using the commands of their own local systems. Projects are now in development to provide Z39.50 support for catalogs on the Internet. SR (Search and Retrieval), ISO Draft International Standard 10162/10163 is the international version of Z39.50.”²

5.6 MIME

*After years of experiments and non-standard, non-interoperating implementations, multimedia mail has yet to become widespread on the Internet or elsewhere, outside of isolated communities. MIME (Multipurpose Internet Mail Extensions), a new standards-track Internet format defined by an Internet Engineering Task Force Working Group, offers a simple standardized way to represent and encode a wide variety of media types, including textual data in non-ASCII character sets, for transmission via Internet mail. MIME extends RFC 822 in a manner that is simple, completely backward-compatible, yet flexible and open to extension. In addition to enhanced functionality for Internet mail, the new mechanism offers the promise of interconnecting X.400 “islands” without the loss of functionality currently found in X.400-to-Internet gateways. This paper describes the general approach and rationale of the new mechanisms for Internet multimedia mail.*³

MIME’s influence has gone far beyond email. The MIME protocol is used by Mosaic and Gopher as a mechanism to communicate various data types. Associating applications with data types enables end users to launch special purpose viewers for special purpose data types. This provides an extensible capability to applications which otherwise would remain closed or difficult to modify.

5.7 HTTP

HTTP the HyperText Transfer Protocol is the native protocol used by WWW.

The following comes from the Internet Draft of the IETF⁴ specification:

*Abstract*⁵

-
1. From the “Internet Terms Glossary” of URLs located at <http://www.vtls.com/glossary.html>.
 2. See <http://www.research.att.com/~wald/pe-doc.txt> “Z39.50 in plain english” by Clifford Lynch for more information.
 3. Nathaniel Borenstein, Internet Multimedia Mail with MIME: Emerging Standards for Interoperability. ULPA A’92 Vancouver, May 1992.
 4. The IETF, Internet Engineering Task Force, is one of the governing bodies of the Internet.

HTTP is a protocol with the lightness and speed necessary for a distributed collaborative hypermedia information system. It is a generic stateless object-oriented protocol, which may be used for many similar tasks such as name servers, and distributed object-oriented systems, by extending the commands, or "methods", used. A feature of HTTP is the negotiation of data representation, allowing systems to be built independently of the development of new advanced representations.

One recent development is S-HTTP, the Commerce Net Secure HTTP Proposal.¹ According to a draft:

Secure HTTP has been designed to enable incorporation of various cryptographic message format standards into Web clients and servers, including, but not limited to, PKCS-7, PEM, and PGP. S-HTTP supports interoperation among a variety of implementations, and is backward compatible with HTTP. S-HTTP aware clients can talk to S-HTTP oblivious servers and vice-versa, although such transactions obviously would not use S-HTTP security features.

S-HTTP does not require client-side public key certificates (or public keys), supporting a symmetric session key operation mode. This is significant because it means that spontaneous private transactions can occur without requiring individual users to have an established public key. While S-HTTP will be able to take advantage of ubiquitous certification infrastructures, its deployment does not require it.

S-HTTP supports end-to-end secure transactions, in contrast with the existing de-facto HTTP authorization mechanisms which require the client to attempt access and be denied before the security mechanism is employed. Clients may be "primed" to initiate a secure transaction (typically using information supplied in an HTML anchor); this may be used to support encryption of fill-out forms, for example. With S-HTTP, no sensitive data need ever be sent over the network in the clear.

6.0 Case Studies

Proof of the viability of network based documents comes through use. The following series of case studies illustrates how a number of organizations are using the WWW both in commercial and noncommercial ways. Each represents a different aspect of the overall user community.

6.1 GNN

One of the most widely known (and oldest) commercial WWW sites is the Global Network Navigator (GNN) run by O'Reilly and Associates, the publisher. GNN offers a

5. See <http://info.cern.ch/hypertext/WWW/Protocols/HTTP/HTTP2.html> for the nitty gritty on HTTP.

1. See <http://www.commerce.net/cgi-bin/textit?/information/standards/drafts/shttp.txt>

great deal of information on the internet, much derived from their book “The Whole Internet Catalog.” In addition, a variety of “special publications” following a magazine style exists on topics such as travel, digital videos and interactive media.



FIGURE 8. GNN home page and one major subdocument

6.2 NIST

At the National Institute of Standards and Technology (NIST), a WWW presents an overview of the entire institute. Up-to-date information about NIST in the news as well as the many programmatic components of NIST are available. The NIST home page provides a centralized collection of pointers to the many other webs at NIST. Most other suborganizations within NIST operate their own web servers.

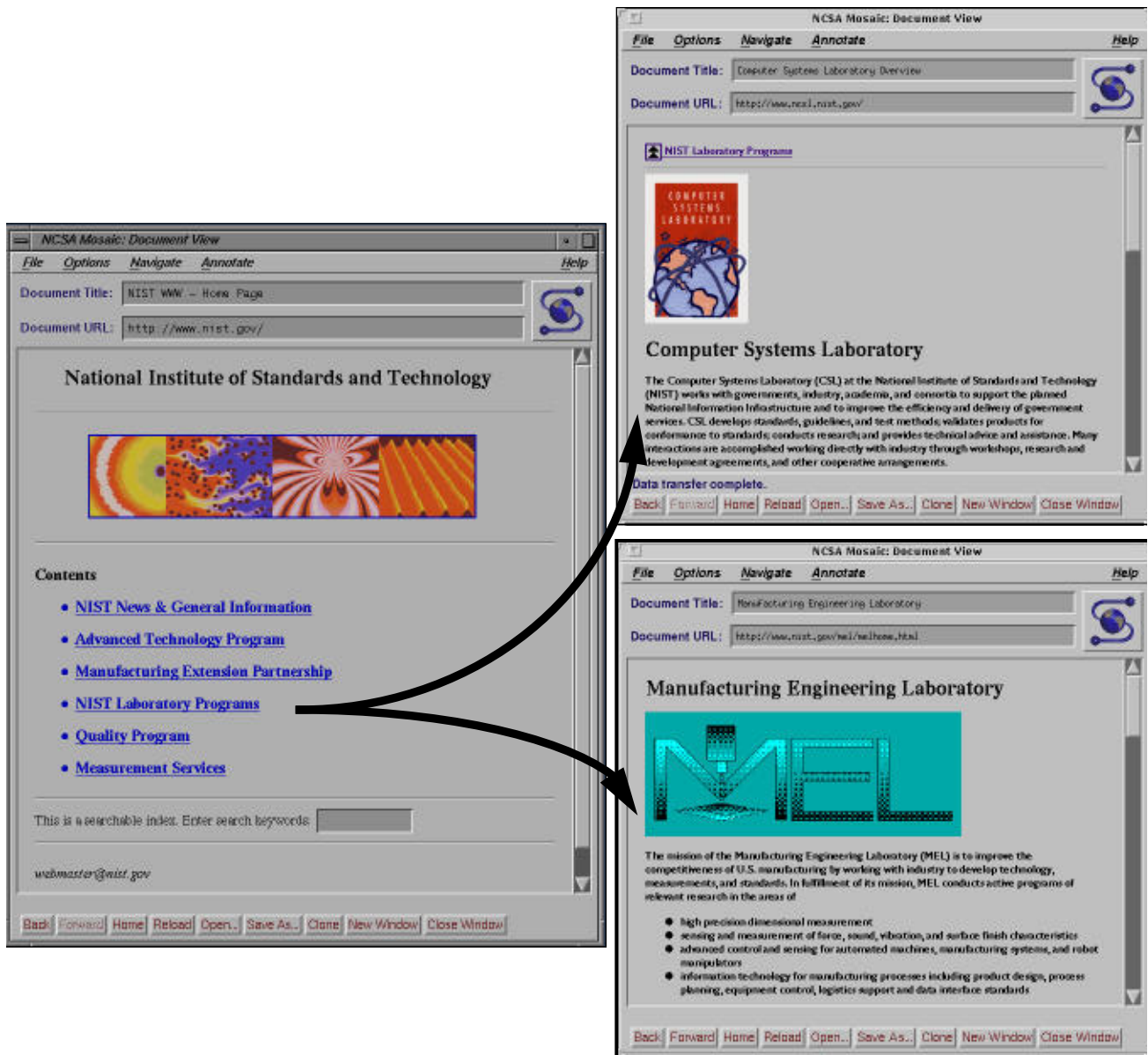


FIGURE 9. NIST main home page and two laboratory home pages.

The development and implementation of a central NIST WWW page took a great deal of internal coordination. The main issues were not technical they were policy and enterprise organizational. The existing policy is to let each laboratory maintain its own web and to have the central web contain a pointer. This solves two problems: each laboratory wants to maintain and control its own information, and the organization responsible for

central computing services does not have the resources to manage the entire institute's many and varied web servers. Each laboratory does however have an information "czar" responsible for ensuring the validity of the laboratories web information.

Often the problems associated with presenting and disseminating information via the WWW are not technical. They are procedural. Issues of responsibility and control often become stumbling blocks along the path toward new information infrastructures. The approach taken by NIST, that of centralizing a small access point and delegating other dissemination points, appears to be a workable solution for a reasonably large organization.

6.3 Library of Congress

The Library of Congress¹ is starting to present a variety of material in different media formats, such as sound, pictures and text on WWW. The potential benefits to the public in terms of increased accessibility is tremendous. According to a description of the project:

The Library of Congress is beginning to use the World Wide Web (WWW) to present information about and materials from its collections over the Internet. This project is in its infancy and will be developing over the next year. Some of the current LC WWW offerings include: photographic and sound collections from the American Memory Project, the African-American Culture and History on-line exhibit, Country Studies from the Federal Research Division, and access to LC MARVEL (LC's gopher-based campus-wide information system) and LOCIS (the Library of Congress Information System). In addition, the Library is developing a Global Electronic Library, which links to WWW meta indexes and search tools, federal government information, and eventually other WWW resources categorized by subject.

1. The Library of Congress is available at the URL: <http://lcweb.loc.gov/>

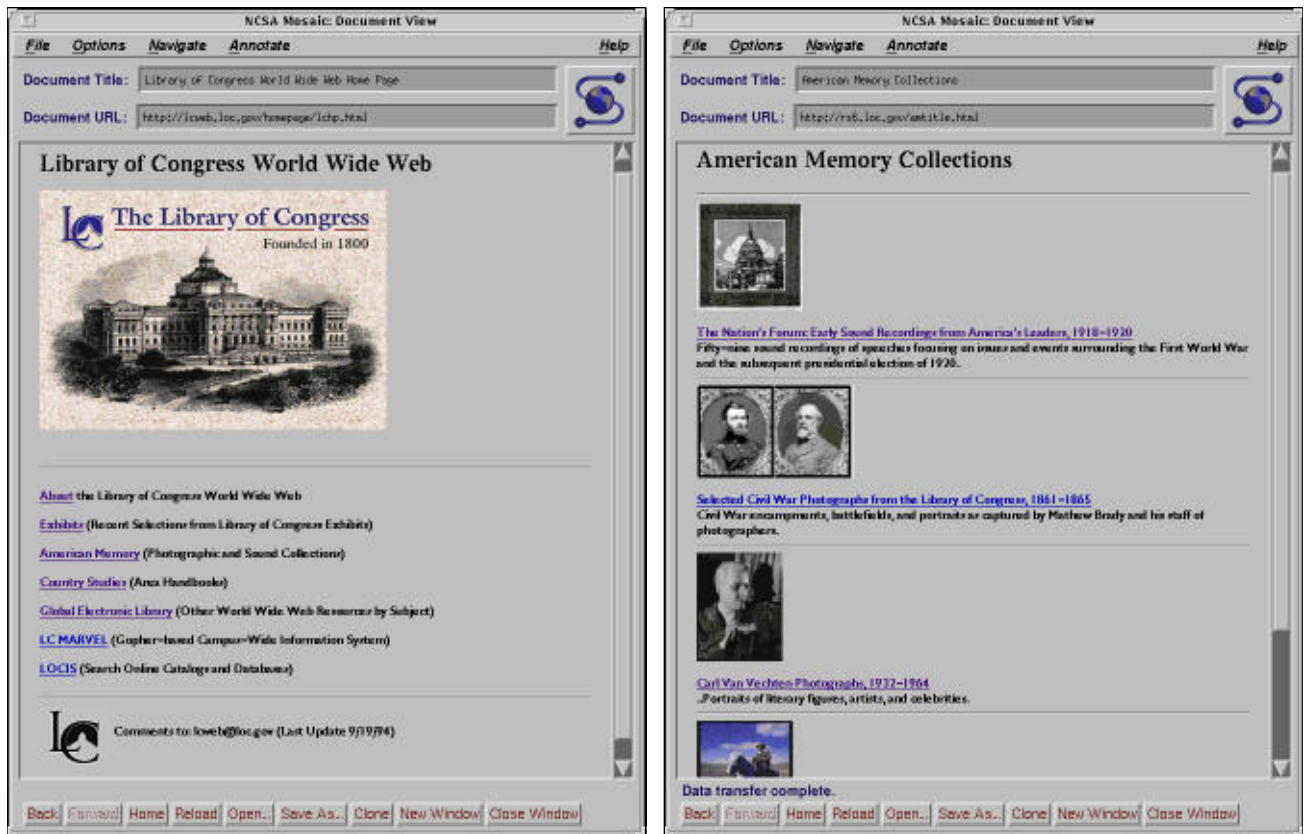


FIGURE 10. The U.S. Library of Congress home page and a photography collection page.

6.4 Geometry Center

Another type of document is exemplified by the work of the Geometry Center in Minnesota. The WWW page points both to “interactive” programs and to multimedia courseware. The following illustrations show the results of interacting with a tiling program which sent back the results as new inline images.

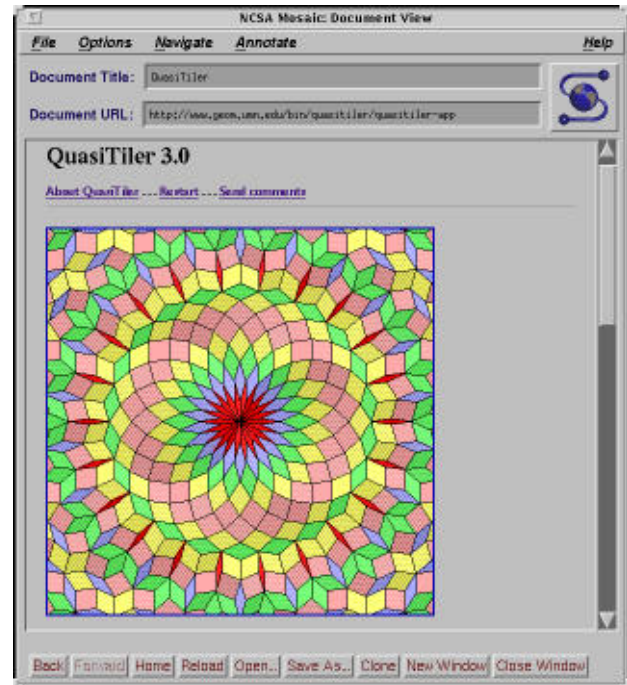
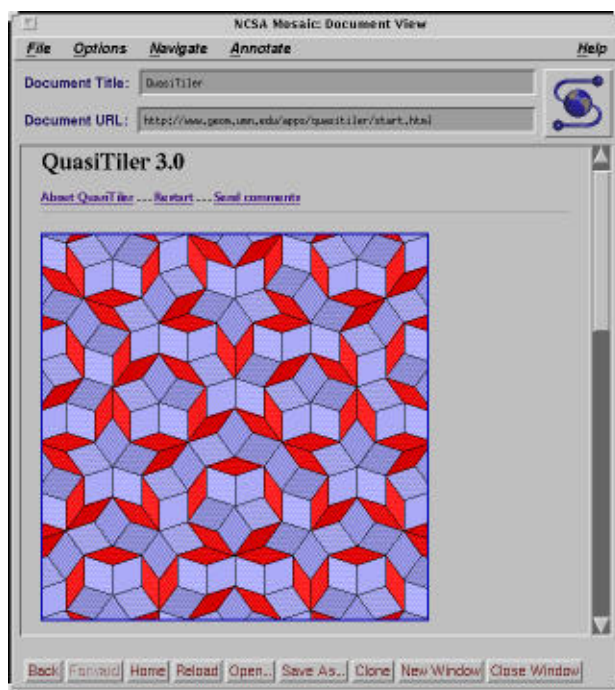
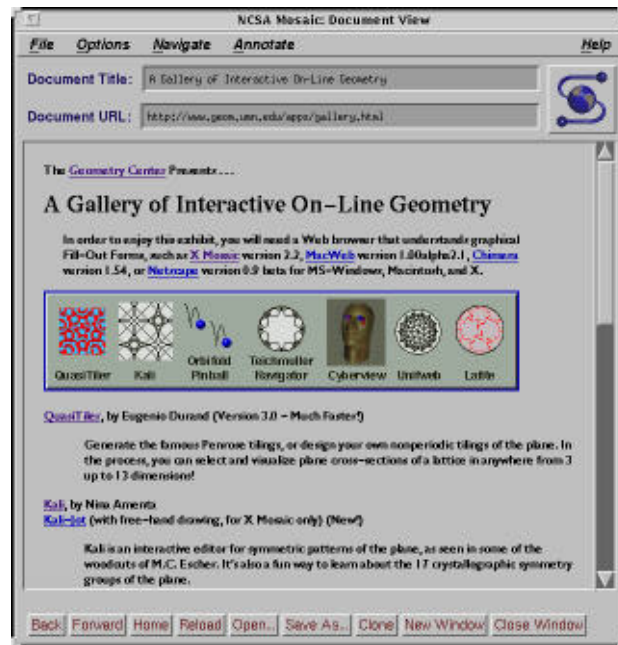


FIGURE 11. The Geometry Center and results of interaction with tiling program via Mosaic

In addition to some pretty pictures, the Geometry Center offers multimedia courseware via WWW. A tutorial and demonstration of four-dimensional geometry and the illustration of a 4D cube, a tesseract, are illustrated in the next figure.

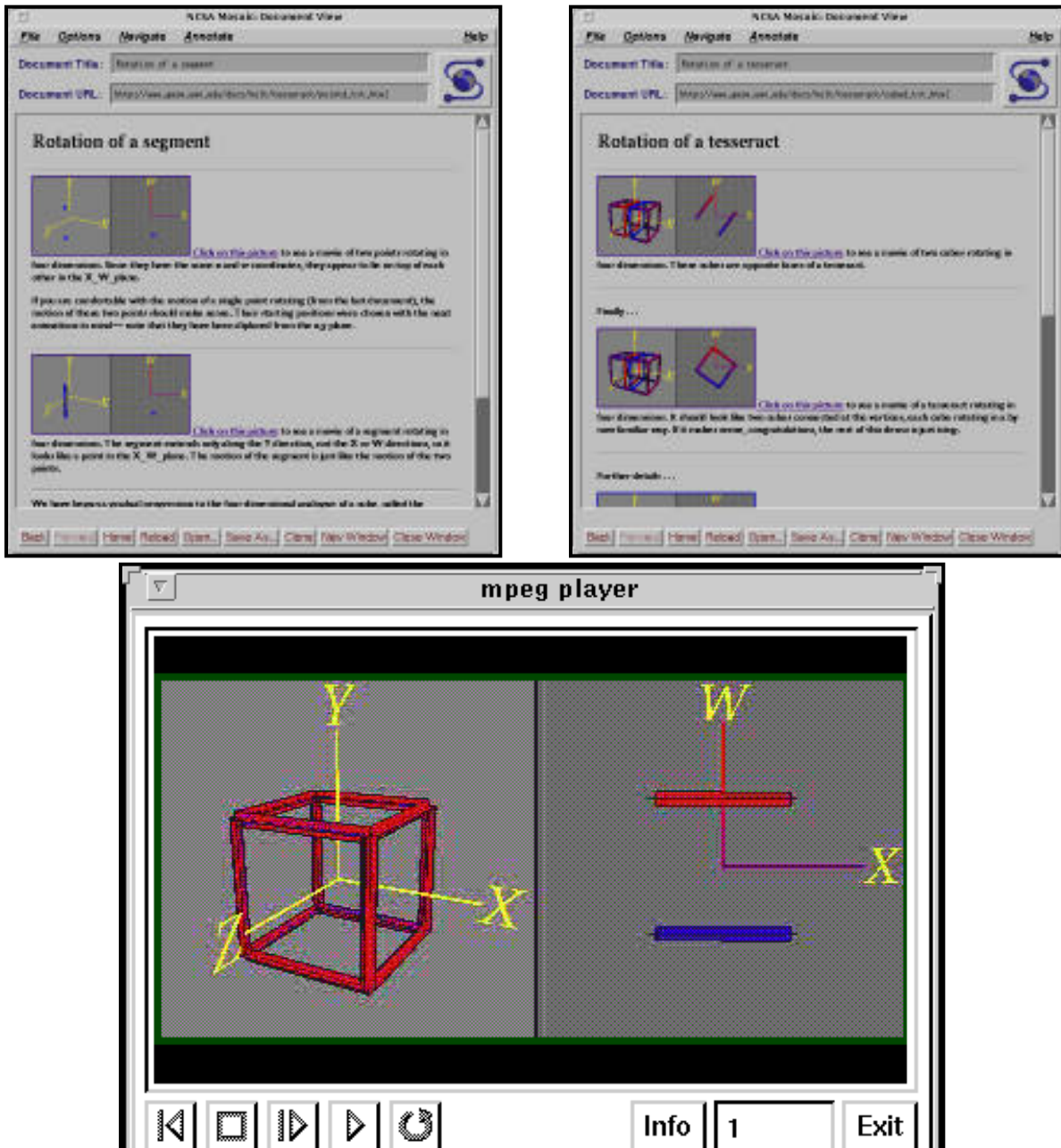


FIGURE 12. Images from an interactive tutorial on 4 dimensional space

Much of the educational benefit of this on-line tutorial derives from the large number of MPEG (Motion Picture Experts Group format) animations. These animations take the user through a step-by-step development process. The motion illustrated by the animations is critical to an understanding of the concepts.

7.0 Conclusions

Networks are providing a new type of dissemination medium. Documents written explicitly for network based distribution can take advantage of new possibilities. Traditional documents written principally for paper dissemination generally retain a quality and page layout advantage over their on-line counterparts. The visual “look” of documents are only now being addressed by on-line systems.

On-line documents also suffer from dependence on a viewer or other operating system capabilities. Once a document is printed it is always “usable.” An on-line document must exist within the context of software and an operating system that also must function correctly. The archival quality of on-line documents is therefore bound to longevity of an unknown party such as the operating system or browser’s vendor. If a vendor ceases to exist or ceases to support a critical feature of the operating system, the on-line document can become useless. This implies that along with the benefits of on-line documents, there are some risks.

The benefits of moving towards on-line network based distribution mechanisms do appear to be compelling enough to risk some problems. Let’s now take a look back at the questions asked at beginning of the chapter.

What is a document in the context of electronic authoring and distribution?

Documents are changing entities. They depend on the environment and are often accompanied by multimedia information.

How can I find the information I want in the vast swamp of information on the net?

A wide variety of tools exists which enable searching. In fact there are too many tools. Archie, WAIS, Veronica and so on are all useful tools. Unfortunately, you must use the tools and learn about sites with rich indices that enable you to find relevant information. In short, roll up your sleeves and sweat a little, there is no magic incantation.

As an author, what can I do to make the information in my documents more usable and useful?

Probably the most useful advise is to structure your information logically. After an initial outline, use the structure-naming tools available in your word processor. Often these are called styles or tags. Documents with meaningful named styles and tags can usually be converted into meaningful logically marked-up documents which can be used and reused. Reuse is a key advantage of properly tagged documents. Information used as a paper document in one instance can be converted for CD-ROM use in another instance and for network browsing in another case.

When all is said and done the content is what usually matters most. However, networked multimedia documents are able to communicate information more effectively than dry information. It is the communication of ideas which ultimately determines the success or failure of any document.

8.0 References

[IIBOX] Internet In A Box will be available in technical bookstore from O'Reilly & Associates and in software stores from Spry.

[ISDN] For a good overview of ISDN and related technologies see "Digital Remote Access" by Jeffrey Fritz in Sept. 1994, Byte.

[KEY] "Pretty Good Privacy" by William Stalling, July 1994 Byte.

[PSI] Performance Systems International, Inc. (PSI) offers a service called InterRamp which provides ISDN access to the Internet. See <http://www.psi.net/> for the details.

[SCHWARTZ] Michael F. Schwartz, Alan Emtage, Brewster Kahle, B. Clifford Neuman, "A Comparison of Internet Resource Discovery Approaches", Computing Systems, Vol. 5, No. 4, Fall 1992.

[SGML] ISO 8879, Standard Generalized Markup Language, SGML.