World-Wide Web: The Information Universe

Tim Berners-Lee, Robert Cailliau, Jean-François Groff, Bernd Pollermann
CERN, 1211 Geneva 23, Switzerland

Abstract

The World-Wide Web (W^3) initiative is a practical project to bring a global information universe into existence using available technology. This article describes the aims, data model, and protocols needed to implement the "web", and compares them with various contemporary systems.

The Dream

Pick up your pen, mouse or favorite pointing device and press it on a reference in this document - perhaps to the author's name, or organization, or some related work. Suppose you are directly presented with the background material - other papers, the author's coordinates, the organization's address and its entire telephone directory. Suppose each of these documents has the same property of being linked to other original documents all over the world. You would have at your fingertips all you need to know about electronic publishing, high-energy physics or for that matter Asian culture. If you are reading this article on paper, you can only dream, but read on.

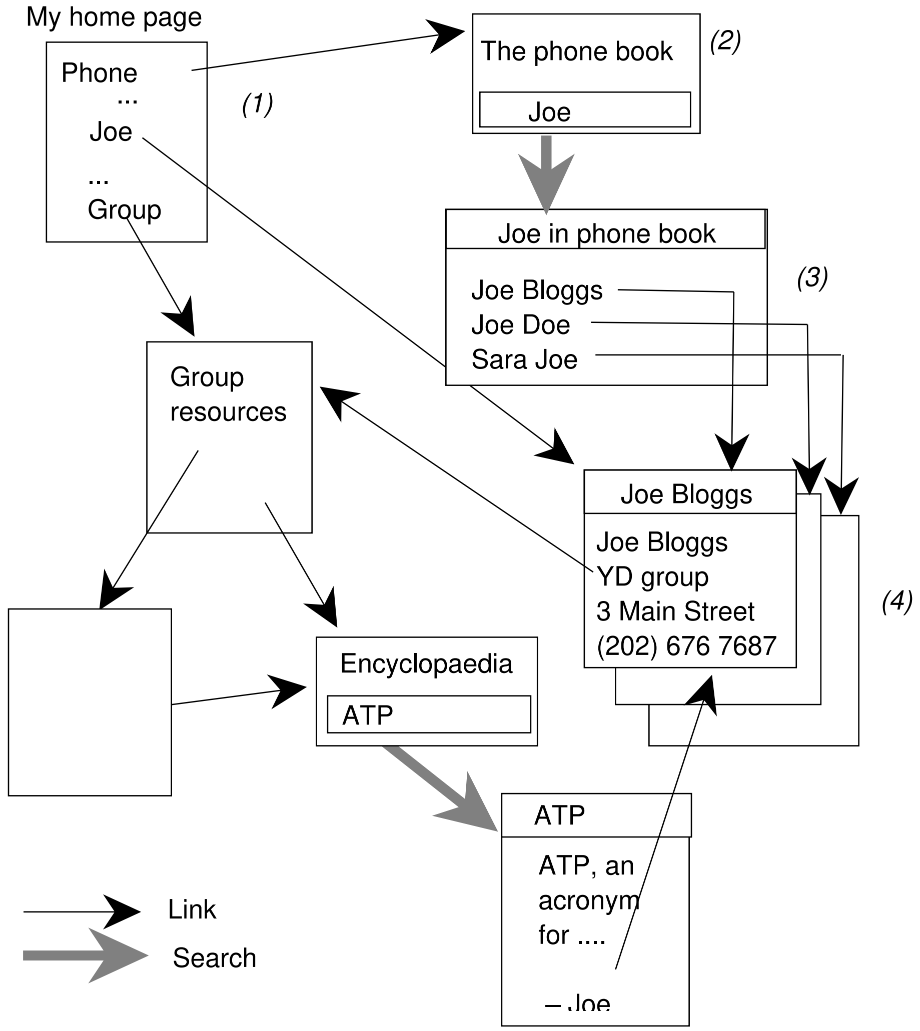
Since Vannevar Bush's article [1], men have dreamed of extending their intellect by making their collective knowledge available to each individual by using machines. Computers give us two practical techniques for the man-knowledge interface. One is hypertext, in which links between pieces of text (or other media) mimic human association of ideas. The other is text retrieval, which allows associations to be deduced from the content of text. In the first case, the reader's operation is typically to click with a mouse (or type in a reference number) - in the second case, it is to supply some words representing that which he desires. The W^3 ideal world allows both operations, and provides access from any browsing platform.

Reality

Existing research projects and commercial products are not far from achieving parts of this dream. The Xanadu system [2] is an ambitious distributed hypertext project. Existing hypertext systems (see for example [3, 4]) tend to be restricted to the local or distributed file system, and often are developed with a limited set of platforms in mind. Contemporary information retrieval and access systems such as Alex [5], Gopher [6], Prospero [7] and WAIS [8] cover a wide area without the hypertext functionality. Merging the techniques of hypertext, information retrieval, and wide-area networking produces the W^3 model. This poses specific requirements on document naming schemes, protocols, and data representation.

The W3 data model

The W^3 model uses both paradigms of hypertext link and text search in a complementary fashion, as neither can replace the functionality of the other. Figure 1 shows how a personalized web of information is built from these operators:



The W3 model involves hypertext links and index searches. The reader starts at the home page (1), and quickly uses his own links, group-wide or public links to find resources. Indexes such as the phone book (2) are represented as documents with the possibility of inputting search words. The result is a virtual hypertext document (3) which points to the documents found (4).

Fig. 1: A web of links and indexes

Features to note are:-

- Information need only be represented once, as a reference may be made instead of making a copy;

- Links allow the topology of the information to evolve, so modeling the state of human knowledge at any time without constraint;

- The web stretches seamlessly from small personal notes on the local workstation to large databases on other continents;

- Indexes are documents, and so may themselves be found by searches, and/or following links. An index is represented to the user by a "cover page" which describes the data indexed and the properties of the search engine.

- The documents in the web do not have to exist as files: they can be "virtual" documents generated by a server in response to a query or document name. They can therefore represent views of databases, or snapshots of changing data (such as the weather forecast, financial information, etc).

A pleasing, and useful, aspect is that almost all existing information systems can be represented in terms of the W^3 model. A menu becomes a page of hypertext, with each element linked to a different destination. The same is true of a directory, whether part of a hierarchical or cross-linked system. The notion of many named indexes within the web allows a given search engine and database to be visible with several different addresses, each representing different options for the search algorithm. For example, the index `/library/books/ti+au/substring` may give a title and author search, whereas `/library/books/text/exact` may give an exact-word full-text search. Addresses are discussed in more detail below.

Publishing

From the information provider's point of view, existing information systems may be "published" as part of the web simply by giving access to the data through a small server program. The data itself, and the software and human procedures which manage it, are left entirely in place. This approach has allowed, for example, a mainframe-based document storage and index system to be opened up to access from all platforms in the organization. To see how this is done requires a brief overview of the W^3 architecture.

W^3 Architecture

Hypertext and text retrieval systems have been available for many years, and a valid question is why a global system has not already come into existence. Traditional answers to this question are the lack of

- A common naming scheme for documents
- Common network access protocols
- Common data formats for hypertext

Most research in hypertext systems (the Xanadu project excepted) has focussed on the user interface and authoring questions, rather than the questions of wide-area and long-term distribution. These architectures have assumed that users share a common application program running on computers (often of the same type) which share a common file system. The W^3 architecture must cope with a widely distributed heterogeneous set of computers running different applications which use different preferred data formats. This requires a client-server model. The client has the responsibility for resolving a document address into a document using its repertoire of network protocols. The server provides data in a simple hypertext or plain text form, or, by negotiation with the client, in any other data format.

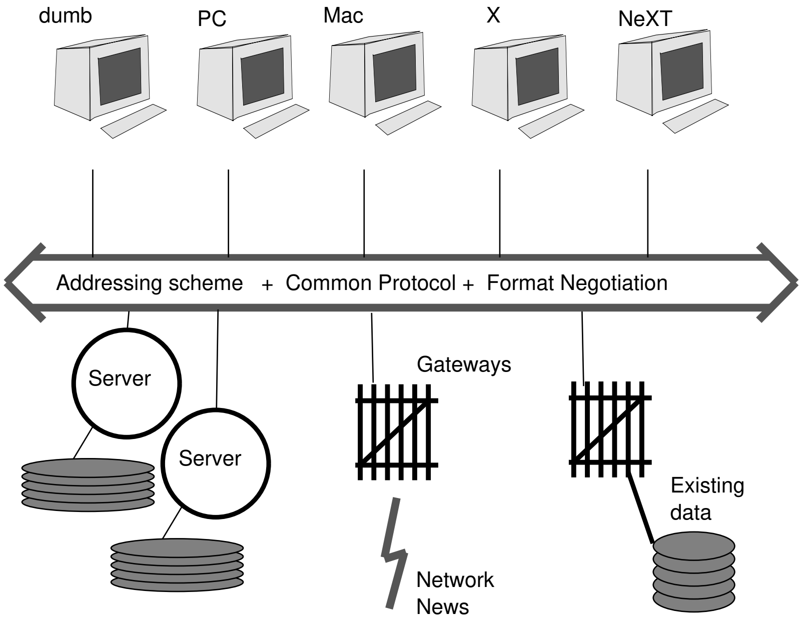


Fig 2. The W^3 architecture in outline.

It may be more difficult initially to develop a generic hypertext browser than a specific front-end for a particular information system. However, the de-coupling of the client and server programs by the "information bus" pays off as more clients and servers are plugged in and universal readership is achieved. Writing a server for new data is generally a simple task because it requires no human interface programming.

Document Naming

The fulcrum on which the document universe rests is the scheme for naming documents. A document name provides a method for the client to find the server, and for the server to find the document. In the W^3 model, a name can also specify a part of the document to be selected by the displaying application.

Although a document name is normally hidden in the hypertext syntax transferred over the link, in practice it must sometimes be referred to by people, and passed through applications (such as mail) which are not yet hypertextaware. It must therefore ideally be composed of printable characters, and manageably short.

Any lasting reference to a document must be a logical name rather than a physical address. That is, it should refer to a document's registration with some "publishing" organization rather than any physical location, so that its location may later be moved. The client is therefore prepared to follow several stages of translation by name servers before finding a final document server. Similarly, a document name should not contain any information which is transitory such as the particular formats available for a document, or its length, for example.

The W^3 naming scheme fulfills these requirements, but is otherwise open to the addition of new protocols as technology evolves. For this purpose a prefix is used to identify the protocol (and therefore naming scheme) to be used. Clients which do not have that protocol in their repertoire refer to a gateway for translation.

Protocols

The W^3 clients are built on a common core of networking code for information access. This core provides access using widely deployed internet protocols such as

- File Transfer Protocol – FTP [9]
- Network News Transfer Protocol - NNTP [10]
- Access to mounted file systems.

A new search and retrieve protocol was found necessary, known as HTTP. Faster than FTP for document retrieval, this also allows index search. HTTP is similar in implementation to the internet protocols above, and similar in functionality to the WAIS protocol. Some differences are discussed below.

Document Formats

The Dexter data model of hypertext [11] provided a conceptual model for hypertext systems, and the HyTime standard [12] formalizes hypertext at a high level. The W^3 project defines a concrete syntax in the SGML style for basic hypertext as used for menus, search results, and on-line hypertext documentation. Every W^3 browsing application is able to parse this simple format *(see Fig. 3)*.

In the pilot phase of the project, this format was all that was required, but in the second phase, format negotiation between client and server will allow the exchange of information in any medium using any mutually acceptable representation.

WAIS and the Web

From the point of view of the W^3 dream, the WAIS protocol represents a significant advance on the search and retrieve (SR) protocol standard Z39.50/ISO-10163, by being stateless, and introducing a persistent name. The document names used are local to the containing database, but these names may be appended to the database name and host address to form a universal W^3

address. In this way, WAIS indexes and servers can be represented in the web. A gateway program, running at CERN and available for general use, provides this mapping. The WAIS model uses separate “source” files to describe indexes. The WAIS-W^3 gateway keeps caches of these files, using them to build descriptive “cover pages” for indexes.

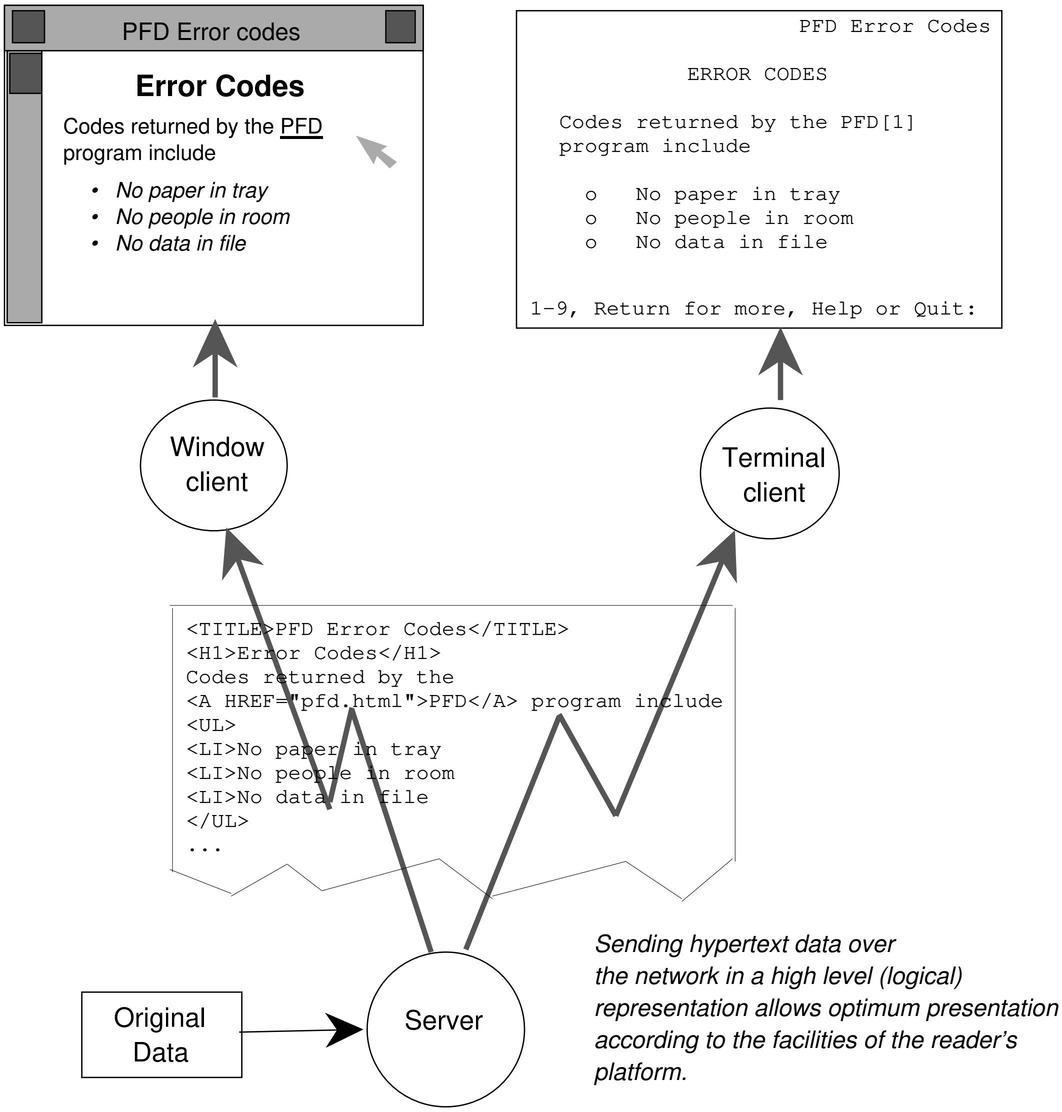


Fig. 3: A schmatic illustration of the encoding of hypertext data. The link is represented in the winodw by underlining, on the terminal by a reference number.

The current WAIS model requires that the results of a search point to documents available from the same server. That is, the same server is responsible for indexing and actually providing the data. In the W^3 world this restriction does not exist. A practical advantage with this approach is that, as Yeong points out

[13], a large multimedia document may be most efficiently retrieved from a different host and using a different protocol to that used for the original query. Futhermore, as on-line information proliferates, an important function is that of "third party" reviewers, indexers and overview writers who refer to data they do not actually hold. It is expected that these services will be a key to the control of the information explosion, and a valuable asset to the community.

A W^3 user builds a personalized web of information by making links from his own notebook into the web. He can make a link to the result of performing a search, such that next time he follows the link the search is reevaluated. This is the equivalent of storing a WAIS "question" - there is a good mapping between the models. The W^3 clients do not currently support relevance feedback although it is not alien to the model.

There are two occasions when hypertext would particularly enhance the WAIS model. Firstly, users often would like to be able to browse through available WAIS indexes. WAIS and W^3 both regard indexes as documents, and therefore allow them to be found using the same techniques as for documents. In fact, the WAIS-W^3 gateway allows a W^3 hypertext overview to be made with pointers to WAIS indexes. Secondly, when one has found a piece of text, WAIS delivers just that part of a file which has been found. Very often one would like links to surrounding information in the same database.

The popularity of WAIS has been a great boost to the world of online information. Its integration with universal naming and hypertext is to be greatly encouraged.

Menu systems and The Web

The Alex[5], Internet Gopher[6] and Prospero[7] systems each use the directory and file (or menu and document) model to implement a global information system. These map into the web very naturally, as each directory (menu) is represented by a list of text elements linked to other directories or files (documents). These systems are very comfortable for readers who are used to hierarchical file systems, for whom directories are an established concept. Even when the structure is in fact cross-linked, the reader feels at home as he regards it as a tree structure. Furthermore, for the information provider such systems are easy to build by cross-linking existing file systems.

An example of mapping a menu system onto the web is made by the W^3 client software which incorporates the simple Gopher protocol, and therefore allows links into the Gopher system. The easy start-up of these systems has made them fairly popular. It is true that a menu is necessarily a more restricting medium of communication than general hypertext: a page of hypertext can convey more information to the reader about the choices to be followed, by using more flexible formatting. Hypertext allows menus of links to lead to nodes with progressively greater textual content. However, the restricted world of plain text and menus, with its ease of publication, is adequate for many information providers.

Similarly, W^3 clients also have built-in ability to browse the world of anonymous FTP archives, and a gateway provides access to Digital™'s

VMS™/Help information.

X.500 and the Web

The x.500 standard for name servers provides a useful tool for long-term naming of documents. Initially intended for coordinates of people and organizations, to be used for documents it needs extensions similar to (though simpler than) those proposed for example by Yeong [14] . The chief attribute of a document for W^3 purposes is the W^3 physical address. Once access to x.500 name servers is widely available, "User Friendly Names" will form an appropriate W^3 document name format for logical addresses.

Experience with the W^3 pilot project

The first client software written to the W^3 requirements ran on the NeXT machine using the NeXTStep™ graphic user interface tools. This hypertext browser/editor demonstrated the ease of use of a window-based hypertext interface to global information. It also allowed an overview hypertext database to be built, to point to data on the web by subject or organization. The second client written was a line-mode browser for character-mode terminals. Being portable to almost any machine, it assures universal readability of all published documents. Hypertext documentation was put on-line, and gateways were set up into various existing information systems.

Enthusiastic users of the browsing software particularly appreciated the consistent user interface for all types of data. Reading news articles as hypertext was a good example: the same user interface is provided, and references between articles, and between articles and the news groups in which they are published, are all consistently represented as links.

It became evident that both hypertext links and text search are important parts of the model. A typical information hunt will start from a default hypertext page by following links to an index. A search of that index may return the required data, or some more links may be followed. Sometimes a further index may be found, and that searched, and so on. When the user of a hypertext editor has found what he wants (no matter how remote), he can make a new link to it from his home page so that he can find it again later almost instantly. This is generally preferable to making a copy which may soon be out of date.

The Future

The success of the pilot project prompted further development of W^3-compliant software and information. Current client projects within various organizations include three X11-based browsers and a Macintosh browser. Various server gateways to other information systems have been produced, and the total amount of information available on the web is becoming very significant, especially as it includes all anonymous FTP archives, WAIS servers and Gopher servers as well as specific W^3 servers. We notice that the functions of each of these servers could be provided by a W^3 server, and so look forward to a single protocol which can be used by the whole community.

The Archie project [15] provides an index into the internet archives and is an excellent example of a service which we hope to make available in the web. We can imagine such indexing being extended to cover other forms of data. W^3 provides a basic infrastructure for information access. All kinds of indexing, searching, filtering and analysis tools could usefully be built using the generic w3 access mechnaism, and so be applied to all the various domains of data. Their results could then be made available on the web. Many possible research projects in hypertext are enabled by the existence of a very large linked information base.

Meanwhile, the W^3 team at CERN and collaborators worldwide invite any information suppliers to join the web, contributing information or software. Detailed information about W^3 protocols and data formats, etc, is available from our W^3 server. The crudest way to access this is by telnet to info.cern.ch. A better way is to run browser software (available by anonymous FTP from the same host) on your local machine. If you use a window-oriented browser, then you will be able to read articles like this on your screen. When you do, pick up your pen, mouse or favorite pointing device and press it on a reference in this document... the dream is coming true.

REFERENCES

[1] Bush, Vannevar, "As We May Think", *The Atlantic Monthly*, July 1945

[2] Nelson, Theodor H., *Literary Machines* version 90.1, Mindfull press 1990.

[3] "Beyond Hypertext: The DECWindows Hyperenvironment Vision", Digital Equipment Corporation, Maynard, MA., 1990

[4] Kahn, Paul and Normal Meyrowitz. "Guide, HyperCard, and Intermedia: A Comparison of Hypertext/Hypermedia Systems", *IRIS Technical Report* 88-7. Brown University, Providence RI, 1988.

[5] Cate, Vincent, Carnegie-Mellon Univerity, private communication.

[6] Alberti et.al. "Notes on the Internet Gopher Protocol" Univeristy of Minnesota, December 1991.

[7] Neuman, Clifford B., "The Prospero File System: User's manual". Department of Computer Science and Engineering, University of Washington.

[8] Kahle, B., et. al., "WAIS Interface Prototype Functional Specification", Thinking Machines Corporation, April 1990

[9] Postel, J. and Reynolds, J. "File Transfer Protocol (FTP)", Internet RFC-959, October 1985.

[10] Kantor, B., and Lapsley, P., "A proposed standard for the stream-based transmission of news", Internet RFC-977, February 1986

[11] Halasz, F. & Schwartz,M., "The Dexter Hypertext reference Model", *Proceedings of the Hypertext Standardization Workshop January 16-18, 1990*, National Institute of Standards and Technology.

[12] GoldFarb, Charles F., *Information Technology – Hypermedia/Time-based Structuring Language (HyTime),* ISO/IEC CD 10744 (Draft).

[13] Yeong, W., "Towards Networked Information Retrieval", *Technical report 91-06-25-01*, Performance Systems International, Inc.

[14] Yeong, W., P.S.I., "Representing Public Archives in the Directory", *Internet Draft*, November 1991.

[15] Emtage, A and Deutch, P, "archie – and Electronic Directory Service for the Internet", to be presented to the 1992 *usenix* conference.