The World-Wide Web

Tim Berners-Lee, Robert Cailliau, Ari Luotonen, Henrik Nielsen, Arthur Secret, CERN, 1211 Geneva 23, Switzerland.

Notes to editors/publishers: changes to this paper

[We traditionally name as authors those involved in the project at the time, so the author list is different from that of the Inet'93 paper.

Much of the "state of the Web" material has had to be updated. I have included some modifications as suggested by reviewers

Figure 1 is as in the original paper. We will find some more screen dump examples, possibly for use in a collage.

Figure 2 is a postscript transparency I have in color, in a separate file.

I have added a graph (Figure 3) of growth as suggested by a reviewer. As I understand you are going to retype the whole thing, I have put it is as large as possible in black and white for camera reduction. I can send you the table from which it was prepared if you want to regenerate it on your own system.

I have been nominally on vacation this last week (hence some time to work on the paper!), but will be back at CERN next week. Phone +41(22)767 3755, fax +41(22)767 7155 at CERN, or +33-50.41.46.94 at home if that's easier for you. Email timbl@info.cern.ch]

Abstract

The World-Wide Web (W3) is a way of viewing all the on-line information available on the Internet as a seamless, browsable continuum. Using hypertext jumps and searches, the user navigates through an information world partly hand-authored, partly computer-generated from existing databases and information systems. The Web today incorporates all information from information systems such as Gopher as WAIS, as well as sophisticated multimedia and hypertext information from many organizations. W3 clients provides comprehensive

point-and-click network access tools, while W3 servers provide an efficient, friendly method of providing data to users. This paper answers some commonly asked questions about W3, comparing W3 with other systems, and about recent developments and future directions for global hypermedia.

1. What is the World-Wide Web?

W3 was developed to be a pool of human knowledge, which would allow collaborators in remote sites to share their ideas and all aspects of a common project. Physicists and engineers at CERN collaborate with many other institutes to build the software and hardware for high-energy physics research. The idea of the Web was prompted by positive experience of a small home-brew personal hypertext system used for keeping track of personal information on a distributed project. The Web was designed such that if it was used independently for two projects, and later relationships were found between the projects, then no major or centralised changes would have to be made, but the information could smoothly reshape to represent the new state of knowledge. This property of scaling has allowed the Web to expand rapidly from its origins at CERN across the Internet irrespective of boundaries of nations or disciplines.

If you haven't come across the Web before, the best way to find out about it is to try it. An appendix to this paper gives some recipes for getting hold of W3 clients. Given one of these, you will quickly find out all you need to know, and much more. For hard copy to read on the plane, or if you don't have Internet access from your desk top machine, refer to our paper in "Electronic Networking" [W3] for an overview of the project, material which we will not repeat but will summarize here.

A W3 "client" program runs on your computer. When it starts, it displays an object, normally a document with text and possibly images. Some of the phrases and images are highlighted: in blue, or boxed, or perhaps numbered, depending on what sort of a display you have and how your preferences have been set. Clicking the mouse on the highlighted area (“anchor”) causes the client program to retrieve another object from some other computer, a "server". The retrieved object is normally again hypertext, and so the process of navigation continues.

Figure 1. A view of NCSA's Mosaic W3 browser for X. Here we are looking at part of the "Virtual Library", one list of Internet resources by subject. The document title and address (labelled "URL") are displayed at the top. Linked text is underlined here, although color is normally used when available. The "Search keyword" box at the bottom is enabled when the server allows for a search at this point. Buttons provide the usual navigation aids, and a "hot list" of pointers to book-marked documents is saved between sessions.

When viewing some documents, the reader can request a search, by typing in plain text (or complex commands) to send to the server, rather than following a link. In either case, the client sends a request off to the server, often a completely different machine in some other part of the world, and within (typically) a second, the related information, hypertext again, plain text or multimedia, is presented. This is done repeatedly, and by a sequence of selections and searches one can find anything which is "out there". Some important things to note are:

- Whatever type of server, the user interface is the same, so users do not need to understand the differences between the many protocols in common use. Before W3, access to networked information typically involved knowledge of many different access "recipes" for different systems, and a different command language for each. The model of hypertext with text input has proved sufficiently powerful to express all the user interfaces, while being sufficiently simple to require no training for a computer user.

- Links can point to anything which can be displayed, including search result lists. (When a query is applied to an object, the resulting object has an address, defined to be the address of the queried object concatenated with the text of the query. As the result object has an address, one can make links to it. Following the link later leads to a re-evaluation of the query.)

- While menus and directories are available, the extra option of hypertext provides a more powerful communication tool. In simple cases, the server program can generate a hypertext view representing (for example) the directory structure of an existing file store. This allows existing data to be put "on the Web" without further human effort.

- There is a very extendible system for introducing new formats for multimedia data;

- There are many W3 client programs. As hypertext information is transmitted on the network in logical (mark-up) form, each client can interpret this in a way natural for the given platform, making optimal use of fonts, colors, and other human interface resources available on that platform.

II. What does W3 define?

W3 has come to stand for a number of things, which should be distinguished. These include

- The idea of a boundless information world in which all items have a reference by which they can be retrieved;

- The address system [URI] which the project implemented to make this world possible, despite many different protocols;

- A network protocol [HTTP] used by native W3 servers giving performance and features not otherwise available;

- A mark-up language [HTML] which every W3 client is required to understand, and is used for the transmission of basic things such as text, menus and simple on-line help information across the net;

- The body of data available on the Internet using all or some of the above.

The client-server architecture of the web is illustrated in figure 2.

Figure 2. The World-Wide Web client-server architecture. For published information to be universally available, W3 relies on a common addressing syntax, a set of common protocols, and negotiation of data formats.

Universal Resource Identifiers

Universal Resource Identifiers[1] are the strings used as addresses of objects (menus, documents, images, etc.) on the Web. For example, the URI of the "welcome" page for the CERN server is

`http://info.cern.ch/hypertext/WWW/TheProject.html`

URIs are "Universal" in that they encode members of the universal set of network addresses. For a new network protocol which has some concept of object, one can form an address for any object as the set of protocol parameters necessary to access the object. If these parameters are encoded into a concise string, with a prefix to identify the protocol and encoding, one has a new URI scheme. The same can be done for names of objects in a given name space.

The prefix "http" in the example above indicates the address space, and defines the interpretation of the rest of the string. The HTTP protocol is to be used, so the string contains the address of the server to be contacted, and a substring to be passed to the server. Different protocols use different syntaxes, but there is a small amount of common syntax. For example, the common URI syntax reserves the "/" as a way of representing a hierarchical space, and "?" as a separator between the address of an object and a query operation applied to it. As these forms recur in several information systems, to allow expression of them in the common syntax allows the features to be retained in the common model, where appropriate. Hierarchical forms are useful for hypertext, where one "work" may be split up into many interlinked documents. Relative names exploit the hierarchical structure and allow links to made within the work independent of the higher parts of the URI such as the server name.

URI syntax allows objects to be addressed not only using HTTP, but also using the other common networked information protocols in use today ([FTP], [NNTP], [Gopher], and [WAIS]), and will allow extension when new protocols become are developed.

[1] The Internet Engineering Task Force (IETF) is currently defining a similar and derived syntax known as a Uniform Resource Locator (URL). As this work is not complete, and there is no guarantee that URLs will have the same syntax or properties as URIs, we use the term URI here to avoid confusion.

URIs are central to the W3 architecture. The fact that it is easy to address an object anywhere on the Internet is essential for the system to scale, and for the information space to be independent of the network and server topology.

Hypertext Transfer Protocol

Perhaps misnamed, rather than being a protocol for transferring hypertext, HTTP is a protocol for transferring information with the efficiency necessary for making hypertext jumps. The data transferred may be plain text, hypertext, images, or anything else.

When a user browses the Web, objects are retrieved in rapid succession from often widely dispersed servers. For small documents, the limitations to the response time stem mainly from the number of round trip delays across the network necessary before the rendition of the object can be started. HTTP is therefore a simple request/response protocol.

HTTP does not only transfer HTML documents. Although HTML comprehension is required of W3 clients, HTTP is used for retrieving documents in an unbounded and extensible set of formats. To achieve this, the client sends a (weighted) list of the formats it can handle, and the server replies with data in any format which it also can produce. This allows proprietary formats to be used between consenting programs in private, without the need for standardization of those formats. This is important both for high-end users who share data in sophisticated forms, and also as a hook for formats which have yet to be invented.

HTTP is an Internet protocol. It is similar in its readable, text-based style to the File Transfer (FTP) and Network News (NNTP) Protocols that have been used to transfer files and news on the Internet for many years. Unlike these protocols, however, HTTP is stateless. (That is, it runs over a TCP connection which is held only for the duration of one operation.) The stateless model is efficient when a link from one object may lead equally well to an object stored on the same server, or to another distant server. The purpose of a reference such as a URI is that it should always refer to the "same" (in some sense) object. This also makes a stateless protocol appropriate, as it returns results based on the URI but irrelevant of any previous operations performed by the client.

The HTTP request from the client starts with an operation code (known as the method, in conformance with object oriented terminology) and the URI of the object. The “GET” method used by all browsers is defined to be idempotent in that its semantics should change the state of the Web (apart from billing for the information transfer and statistics). A “PUT” method is defined for front-end update, and a “POST” method for the attachment of a new document to the Web, or submission of a filled-in form or other object to some processor. Use of PUT and POST is currently in limited, partly due to scarcity of hypertext editors. The extension to other methods is the subject of study.

When objects are transferred over the network, information about them (“meta-information”) is transferred in HTTP headers. The set of headers is an extension of the Multipurpose Internet Mail Extensions [MIME] set. This design decision was taken to open the door to integration of hypermedia mail, news, and information access. Unlike in electronic mail, transfer in binary, and transfer in non standard but mutually agreed document formats is possible. This allows, for example, servers to indicate links from, and titles of, documents (such as bit map images) whose data format does not otherwise include such information.

The convention that unrecognised HTTP headers and parameters are ignored has made it easy to try new ideas on working production servers. This has allowed the protocol definition to evolve in a controlled way by the incorporation of tested ideas.

Hypertext Markup Language (HTML)

Despite the ability of HTTP to negotiate formats, W3 needed a common basic language of interchange for hypertext. HTML is that language, and much of the fabric of the Web is constructed out of it. It was designed to be sufficiently simple so as to be easily produced by both people and programs, but also to adhere to the SGML standard in that a valid HTML document, if attached to SGML declarations including the HTML “DTD”, may be parsed by an SGML parser. HTML is a mark-up language which does not have to be used with HTTP. It can be used in hypertext electronic mail (it is proposed as a format for MIME), news, and anywhere where basic hypertext is needed. It includes simple structure elements, such as several levels of headings, bulleted lists, menus and compact lists, all of which are useful when presenting choices, and in on-line documents.

Under development is a much enriched version of HTML known has HTML+. This includes features for more sophisticated on-line documentation, form templates for the entry of data by users, tables and mathematical formulae. Currently many browsers support a subset of the HTML+ features in addition to the core HTML set.

HTML is defined to be a language of communication, which actually flows over the network. There is no requirement that files are stored in HTML. Servers may store files in other formats, or in variations on HTML which include extra information of local interest only, and then generate HTML on the fly with each request.

W3 and Other Systems

Two other systems, WAIS (from Thinking Machines Corporation an now WAIS, Inc.) and Gopher (from the University of Minnesota), share W3's client-server architecture and a certain amount of its functionality. Table 1 indicates some of the differences.

	WAIS	**Gopher**	**World-Wide Web**
Original target application	Text based information retrieval	Campus-wide information (CWIS)	Collaborative work
Typical objects			
Text	yes	yes	yes
Menus, Graphics	-	yes	yes
Hypertext	-	-	yes
Search functions			
Text search	yes	yes	yes
Relevance feedback	yes	-	-
Reference to other servers	-	yes	yes
Registered servers			
April 1993	113	455	62
April 1993	478*	1410	829

*Table 1. A comparison of three popular network information projects. Registered server figures taken 27 April 1993 and 7 April 1994. WAIS: from TMC directory, number of distinct hosts (*1993: databases). Gopher: from "All the Gophers in the world" register at UMN; W3: from Geographical list kept at CERN. In all cases many more servers exist which are not directly registered, so these are a very rough guide only with no indication of quantity or quality of information at each host*

The WAIS protocol is influenced largely by the z39.50 protocol designed for networking library catalogues. It allows a text-based search, and retrieval following a search. Indexes to be searched are found by searching in a master index. This two-stage search has been demonstrated to be sufficiently powerful to cover the current world of WAIS data. There are however no navigational tools to allow the reader to be shown the available resources, or guided through the data: the reader is parachuted in to a hopefully relevant spot in the information world, but left without context.

Gopher provides a free text search mechanism, but principally uses menus. A menu is a list of titles, from which the user may pick one. While gopher space is in fact a web containing many loops, the menu system gives the user the impression of a tree. The Veronica server provides a master index for gopher space.

The W3 data model is similar to the gopher model, except that menus are generalized to hypertext documents. In both cases, simple file servers generate the menus or hypertext directly from the file structure of a server. The W3 hypertext model gives the program more power to communicate the options available to the reader, as it can include headings, and various forms of list structure for example within the hypertext.

All three systems allow for the provision of graphics, sound and video, although as the WAIS system only has access by text search, text has to be associated with graphics files to allow them to be found.

W3 clients provide access to servers of all types, as a single simple interface to the whole Web is considered very important. Unknown to the user, several protocols are in use behind the scenes. A common code library "libwww" put into the public domain by CERN has promoted this uniformity. Whereas one would not wish to see greater proliferation of protocols, the existence of more than one protocol probably allows for the most rapid progress during this phase in the development of the field. It also allows a certain limited confidence that, if an architecture can encompass older systems and allow transition to current systems, it will, by induction, be able to provide a transition to newer and better ideas as they are invented.

III. Recent W3 developments

This article, like others in this issue, was derived from the paper written in April 1993 for INET'93. Growth of the Web since that time has been so great that this section has had to be completely re-written. There are now 829 rather than 62 registered HTTP servers, and about twice as many client programs available.

The initial prototype W3 client was a "wysiwyg" hypertext browser/editor using NeXTStep. We developed a line mode browser, and were encouraging the developments of a good browser for X workstations. One year ago, NCSA's "Mosaic" W3 browser was in wide use on X workstations. Its easy installation and use was an major reason for the spread of the Web. Today there many browsers available for workstations, Macintosh and IBM/PC compatible machines, and for users with character-based terminals. Of the latter category, "Lynx" from the University of Kansas provides full screen access to the Web for users with character terminals or emulators running on personal computers. As new software is appearing frequently, the reader is advised to check the lists on the Web for those most suitable for his or her needs.

The availability of browsers and the availability of quality information have provoked each other. One available indicator of growth has been Merit Inc.'s count of the traffic of various different protocols across the NSF T3 backbone in the USA *(Fig 3).*

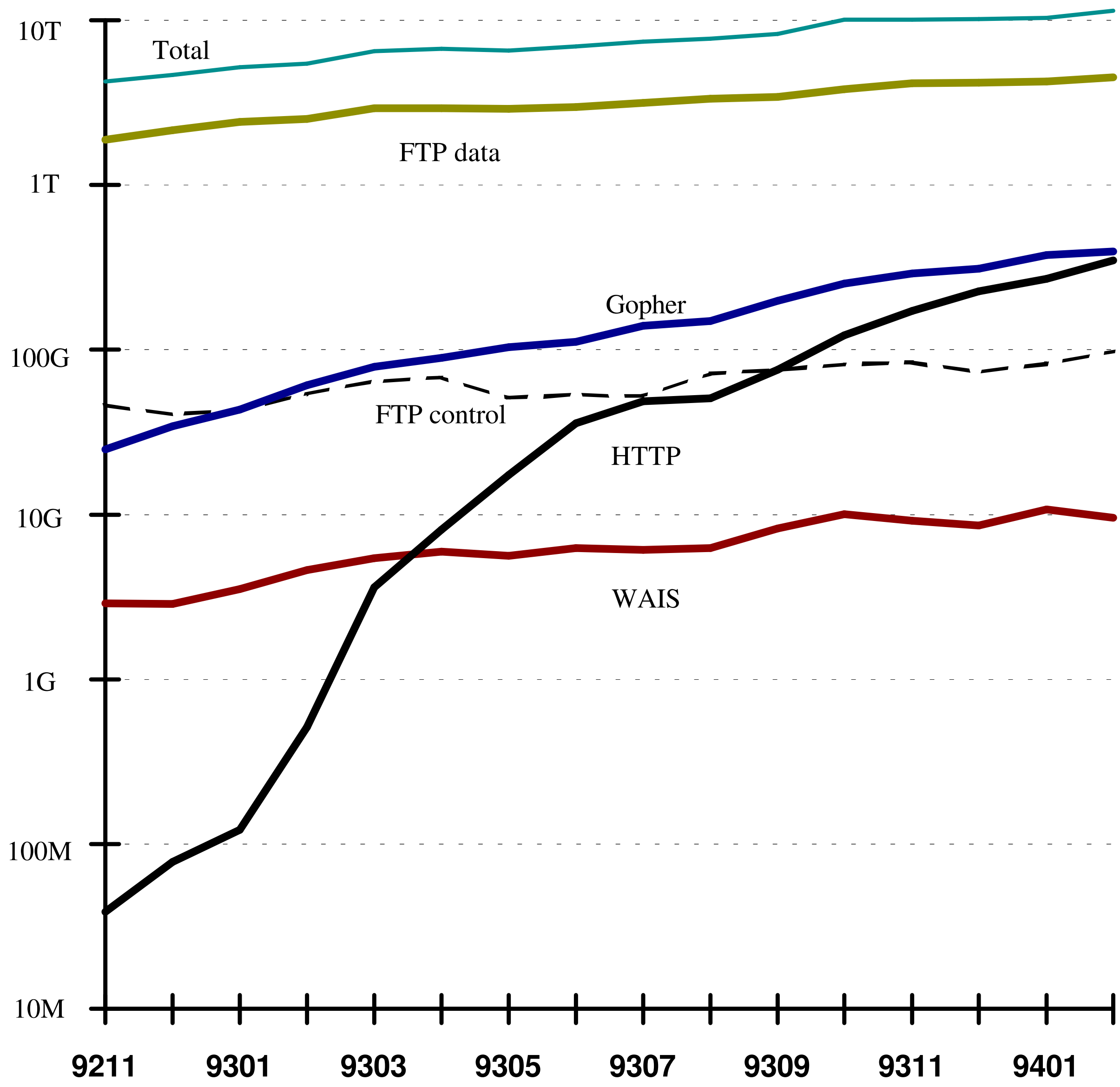

Figure 3.Traffic in bytes per month across the NSF T3 backbone in the USA. File Transfer Protocol (FTP) was traditionally used to access archives of software. FTP uses separate connections for control and data flow. WAIS arose as an interface to text retrieval systems, Gopher protocol with menu style interfaces, and W3's HTTP with hypertext and multimedia. W3 clients handle many protocols to access all these worlds of data as a seamless continuum, but new W3 servers use HTTP by preference. Each vertical division represents a tenfold increase in traffic. The horizontal divisions are months. Data: Merit `<ftp://ftp.merit.edu/statistics/nsf>`

An indicator of the uptake rate of clients is the load on the "info" W3 server at CERN, which more than doubled every 4 months over the two years between April 1991 and 1993.

Information providers have also blossomed. Some of these provide simple overviews of what is available at particular institutes or in particular fields. Others use the power of the W3 model to provide a virtual world of great richness. Examples of servers which use hypertext in interesting ways are the RAL-Durham Particle Database, the Legal Information Institute's hypertexts of several great tomes of American law. Franz Hoesel's hypertext version of the Vatican's Renaissance Culture exhibit at the Library of Congress set an example which was followed by many collections of art, history and other fields. Palo Alto town hall runs a server with everything from building regulations to restaurants. As an example of the increasing use of the Web for commerce, a user-friendly virtual clothing store prompts for one's size, and points to a virtual store which contains only those clothes which are the right size and also in stock.

IV. The future

The W3 initiative sits at the meeting point of many fields of technology. Users put pressure and effort into bringing about the adoption of W3 in new areas. Apart from being a place of communication and learning, and a market place, the web is a show ground for new developments in information technology. Some of the developments which we look forward to in the next few years include

- The implementation of a name service which will allow documents to be referenced by name, independent of their location;

- Hypertext editors allowing non-expert users to make hypertext links to organize published information. This will bring the goal of computer-supported collaboration closer, with front-end update, and annotation.

- More sophisticated document type definitions providing for the needs of commercial publishers of on-line material;

- The development of a common format for hypertext links from two- and three- dimensional images giving more exciting interface possibilities;

- Integration with concurrent editors and other real time features such as teleconferencing and virtual reality;

- Easy-to-use servers for low-end machines to ease publication of information by small groups and individuals;

- Evolution of objects from being principally human-readable documents to contain more machine-oriented semantic information, allowing more sophisticated processing;

- Conventions on the Internet for charging and commercial use to allow direct access to for-profit services.

Conclusion

Hopefully after reading this article you will have an idea of what W3 is, where it fits in with other systems in the field, and where it is going. There is much more to be said, especially about providing information, but this is described on the Web itself. Also in the "Web about the Web" are lists of contributed research and development work and ideas, and pointers to work in progress, so that those interested can work together.

The Web does not yet meet its design goal as being a pool of knowledge which is as easy to update as to read. That level of immediacy of knowledge sharing waits for easy to use hypertext editors to be generally available on most platforms. Most information has in fact passed through publishers or system managers of one sort or another. However, the incredible diversity of information available gives great credit to the creativity and ingenuity of information providers, and points to a very exciting future.

VI. References

FTP	File Transfer Protocol. J. Postel and J. Reynolds, File Transfer Protocol, Internet RFC 959, October 1985.
Gopher	The Internet Gopher. F. Anklesaria et. al., The Internet Gopher Protocol, Internet RFC 1436, March 1993. ftp://ds.internic.net/rfc/rfc1436.txt
HTML	Hypertext Markup Language. T. Berners-Lee and D. Connolly, Hypertext Markup Language, ftp://info.cern.ch/pub/www/doc/html-spec.ps, .txt

HTTP	Hypertext Transfer Protocol. T. Berners-Lee, Hypertext Transfer Protocol, ftp://info.cern.ch/pub/www/doc/http-spec.ps, .txt
MIME	Multipurpose Internet Mail Extensions. N. Borenstein and N. Freed, MIME (Multipurpose Internet Mail Extensions): Mechanisms for Specifying and Describing the Format of Internet Message Bodies", Internet RFC 1341, June 1992
NNTP	Network News Transfer Protocol. Kantor, B & Lapsley, P (1986). A proposed standard for the transmission of news. Internet RFC 977
URI	Universal Resource Identifier. T. Berners-Lee, Universal Resource Identifiers for the World-Wide Web, subitted as an Internet RFC as yet unnumbered. See http://infocern/ch/hypertext/WWW/Addressing/Addressing.html for pointers to information on this area.
WAIS	Wide Area Information Servers. See T. Addyman, WAIS: Strengths, Weaknesses and Opportunities, proceedings of "Information Networking 93", London, May 1993, Meckler, London.
W3	T.J. Berners-Lee, R. Cailliau, J-F Groff, B. Pollermann, CERN, "World-Wide Web: The Information Universe", published in "Electronic Networking: Research, Applications and Policy", Vol. 2 No 1, pp. 52-58 Spring 1992, Meckler Publishing, Westport, CT, USA. See also documents in ftp://info.cern.ch/pub/www/doc and information referenced by http://info.cern.ch/hypertext/WWW/TheProject.html

Appendix

Getting Started

If you have a vt100 terminal, you can try out a full screen interface by telnet to ukanaix.cc.ukans.edu and logging in as www. With any terminal, you can telnet to info.cern.ch for the simplest interface. These browsers are also available in source and in some cases binary form. Details of status and coordinates of about 20 different browsers are available on the Web - just follow a link to World-Wide Web, and select "software available".

The kernel W3 code (a common code library, and basic server and clients) from CERN is in the public domain. (All protocols and specifications are public domain.

NCSA's "Mosaic" browser for W3 is available for X, Mac or PC/Windows by anonymous FTP from ftp.ncsa.uiuc.edu, currently without charge for academic users.

Author Information

Tim Berners-Lee originated the World-Wide Web in 1990, to enable the sharing of knowledge in complex distributed groups of teams. At CERN, the European Particle Physics Laboratory in Geneva, Switzerland, he now coordinates W3 development by collaborating institutes around the world. Originally acquiring his BA in physics from Oxford University in 1976, he immediately moved into the software industry. His experience since has been in text processing, graphics, and communications software, and system design with a number of companies in the United Kingdom. In 1984 Tim moved to CERN where, prior to initiating the W3 project, he developed tools for distributed real-time software. Tim may be reached at timbl@info.cern.ch

Formerly in programming language design and compiler construction, Robert Cailliau has been interested in document production since 1975, when he designed and implemented a widely used document markup and formatting system. He ran CERN's Office Computing Systems group from 87 to 89. He is a long-time user of

Hypercard, and has been working on W3 since 1991, contributing many ideas, and software for the Macintosh. He now coordinates the use of W3 by CERN experiments and other physics institutes. Robert's mail address is cailliau@cernnext.cern.ch.

Ari Luotonen is a member of CERN's technical student program in conjunction with his studies at Tampere University of Technology, Finland. He is developing CERN's "httpd" HTTP server for unix and VMS systems in many areas, including access authorization. Ari 's email address is luotonen@ptsun00.cern.ch.

Henrik Nielsen, of Aalborg University, Denmark, is also a CERN technical student. He is working on the kernel code, with research interests in enhanced networking protocols. His email address is frystyk@dxcern.cern.ch

Arthur Secret wrote the first Gateway giving W3 access to a relational database in 1992, while studying Computer Science at Ecole Internationale des Sciences du Traitement de l'Information in Paris, France, as a CERN's Technical Student. Among other tasks in the CERN WWW team, Arthur currently organizes the cataloguing of new W3 material in the "virtual library".