Computer Networks and ISDN Systems 25 (1992) 454–459
North-Holland

# Networked information services

# The world-wide web

T.J. Berners-Lee, R. Cailliau and J.-F. Groff
*CERN, 1211 Geneva 23, Switzerland*

*Abstract*

Berners-Lee, T.J., R. Cailliau and J.-F. Groff, The world-wide web, Computer Networks and ISDN Systems 25 (1992) 454–459.

This paper describes the World-Wide Web (W3) global information system initiative, its protocols and data formats, and how it is used in practice. It discusses the plethora of different but similar information systems which exist, and how the web unifies them, creating a single information space.
We describe the difficulties of information sharing between colleagues, and the basic W3 model of hypertext and searchable indexes. We list the protocols used by W3 and describe a new simple search and retrieve protocol (HTTP), and the SGML style document encoding used. We summarize the current status of the X11, NeXTStep, dumb terminal and other clients, and of the available server and gateway software.

*Keywords:* global information; hypertext; world-wide web; networked information retrieval; application; browser; server.

## Introduction

This paper covers material presented or elicited by questions at the JENC92 conference. The dream of global hypertext and its coming to fruition with W3 has been described in [1] which also discusses the relationship with other projects in the field. The practicalities of publishing data on the web are outlined in [2], so these aspects will only be summarized here.

## The aim

Much information is available today on the network, but most is not. When an individual enters a new organization, or a new field, it is normally necessary to talk to people, look on bookshelves and nose around for clues about how the place works, what is new, and what he or she needs to know.

When data is available on the net, the average person is not privy to it, but must consult a "guru" who understands the ins and outs of anonymous FTP, telnet, stty, and the command systems of the various information servers.

The aims of the W3 initiative are twofold: firstly to make a single, easy user-interface to all types of information so that all may access it, and secondly to make it so easy to add new information that the quantity and quality of online information will both increase. Already, most information of value exists in some machine-readable form: if we can solve the problems of heterogeneity of platform, data format and access protocol the resulting universe of knowledge will considerably enhance our working together.

## The W3 model

This is described at more length in [1] but is basically as follows. At any time, the user is

Correspondence to: Mr. T.J. Berners-Lee, CERN, 1211 Geneva 23, Switzerland. Tel. (+41) 22 76 73 755, Fax (+41) 22 76 67 7155, E-mail: timbl@info.cern.ch.

Elsevier Science Publishers B.V.

reading a document. Two navigation operations are provided: one operation is to follow a link from a particular piece of text to a related document or part of a document. The other operation is to query a server with a text string. The result of either operation is the display of a new document. The query operation is only provided for certain documents: those which represent a search facility provided by a remote server. The query typically returns a synthesized hypertext list of items, each linked to some document which matches the query.

The power of hypertext links, whether generated automatically or authored by a human being, to represent knowledge in a way easy to follow by the reader cannot be replaced by powerful query languages. Conversely, the power of special-purpose query engines to solve user-posed queries cannot be replaced by a system using only hypertext links. The two operations are both found to be necessary.

One view encompasses all the systems.

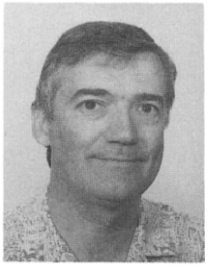
Although a simple model, a significant point in its favour has been its ability to represent almost all existing information systems. This gives us both a great start in putting existing information online, and also confidence that the model is one which will last for future information systems.

(Some information systems include complex search functions which have many fields to fill in. However, when the human–computer interaction is analysed, it can be broken down into a sequence of choices and user input. This leads in general to a simplified user interface and a direct mapping onto the web.)

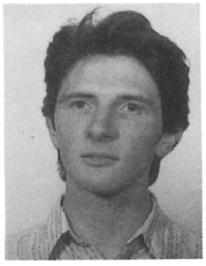
Systems which have been mapped onto the web to date include WAIS, Gopher, VMS(TM)/Help, FTP archives, The "Hyper-G" system from the Technical University of Graz, Gnu TexInfo, unix manual pages, unix "finger", and several proprietary documentation sytems. A W3 client provides a seamless view of all this data using a simple user interface. For the reader, therefore, W3 solves the problems of the over-abundance of different information systems.

An information provider, however, may wonder which management system to use on his server. There is no single recommended W3

**Tim Berners-Lee**, before coming to CERN, worked on, among other things, document production and text processing. He developed his first hypertext system, "Enquire", in 1980 for his own use (although unaware of the existence of the term HyperText). With a background in text processing, real-time software and communications, Tim decided that high energy physics needed a networked hypertext system and CERN was an ideal site for the development of wide-area hypertext ideas. Tim started the WorldWideWeb project at CERN in 1989. He wrote the application on the NeXT along with most of the communications software.



**Robert Cailliau**, who was formerly in programming language design and compiler construction, 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 1987 to 1989. He is a long-time user of Hypercard, which he used to such diverse ends as writing trip reports, games, bookkeeping software, and budget preparation forms. When he is not doing WWW's public relations, Robert is contributing browser software for the Macintosh platform, and analysing the needs of physics experiments for online data access.



**Jean-François Groff** provided useful input in the "design issues". During his stay at CERN as "cooperant", J-F joined the project in September 1991. He wrote the gateway to the VMS Help system, worked on the new modular browser architecture, and helped support and present WWW at all levels. He is now porting the communications code to DECnet in order to set up servers for physics experiments.

server. Fortunately, it is normally true that if he has a relatively organized approach to keeping his data, he can generally adopt new tools without the user being very aware that things are different. For example, he can run a new indexing system, or generate a new browsable hierarchy, on top of his existing data. He should pay close attention to the easy collection or contribution and update of the data, as this is the step which will ensure its quality. No amount of work on the access software can make up for a lack of accuracy or timeliness of the raw data. The W3 project does not restrict the choice of databases for information management; it simply makes whatever exists visible. However, it is hoped that more simple tools for automatically publishing files and mail archives as indexed hypertext will be available in the near future.

It is important to emphasize that although the user model is hypertext, the data published does not have to be prepared in hypertext form. Most data on the web is plain text, accessed by automatically generated hypertext links (following for example the directory tree in which the files are stored) or an automatic indexing system. The prospective information provider need not be frightened by the term "hypertext". It is true, however, that he may in the end wish at least his

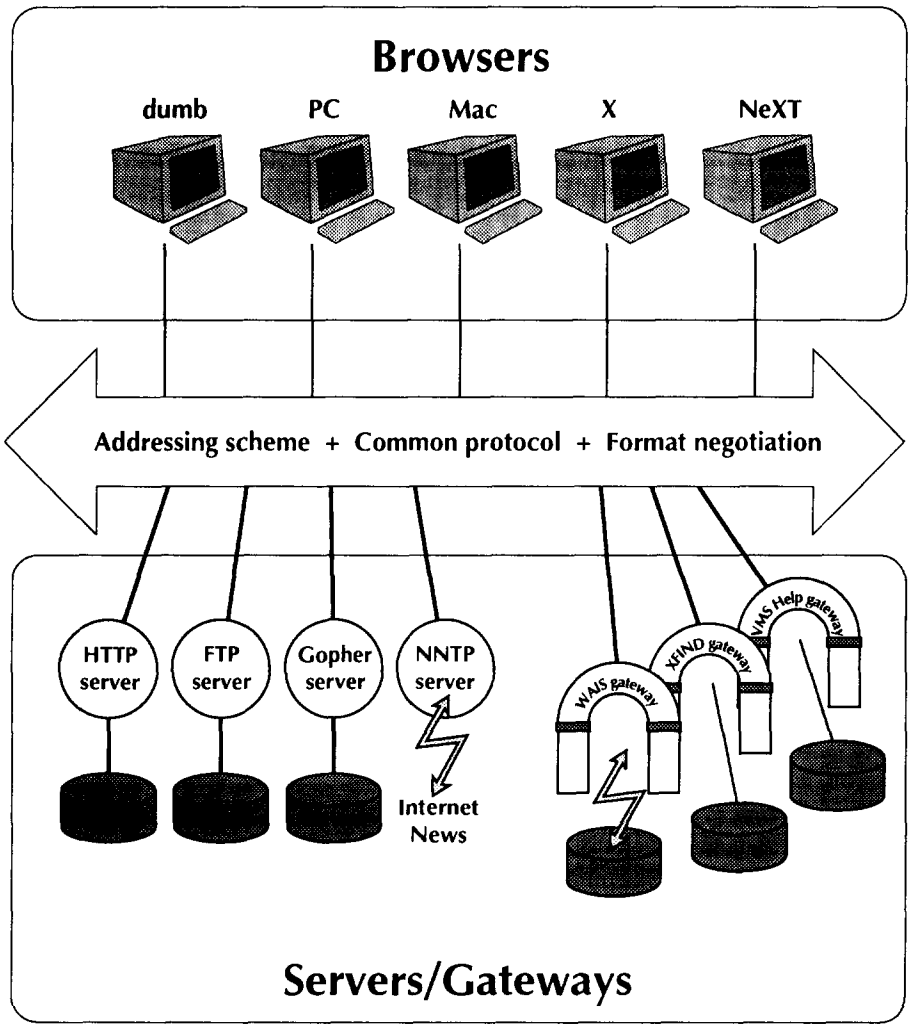



Fig. 1. The world-wide web client-server architecture.

overview document to be hand-written hypertext in order to maximize the impact and communication with the reader.

## Protocols

W3 uses a client-server architecure (Fig. 1), to allow complex presentation facilities to be provided by the client, and powerful search and data manipulation algorithms to be provided at the site of the data by a server. The protocol needed to connect server and client is a simple stateless search and retrieve protocol. In practice the W3 clients all include the ability to use various other protocols, including FTP [3], Gopher [4], local file access, and NNTP [5] for internet news. This gives each W3 client access to several already existing worlds of information. The document addressing scheme allows names to be given to any document, file, directory, newsgroup or article in these systems. This means that a hypertext document may be written or generated which includes links to these objects. Other worlds of information such as that of WAIS servers [6] are made available by gateways which perform the mapping of that world into the web.

Ideally, a protocol used by W3 has the following features:

- Document retrieval by name.
- Index search by name of index plus reader-supplied text.
- Stateless operation. Rapid traversal of links between documents on different servers makes the concept of a session between client and server inappropriate.
- Minimum number of round-trips. As technology advances, processing time may continue to shrink, leaving long-distance round trip delays the dominant factor in response time.
- Pipelining allowing the first part of a document to be displayed (or relayed through a gateway) before the whole document has been transferred. This is easy when using a byte stream oriented protocol.

To achieve this, a simple new protocol, HTTP (Hypertext Transfer Protocol) was defined in the conventional Internet style. This runs over TCP/IP, using one TCP/IP connection per search or retrieve operation.

The initial form of the protocol involves the client sending a simple ASCII request for a document: the command "GET" and the document's name. The response to this is either a hypertext file marked up in SGML, using a specific document type known as "HTML", or a plain text document with an HTML header. In the new version of the protocol (under development) an SGML-formatted request object includes details about the client capabilities. The client capabilities include a weighting, in the form of penalty points for information loss and time taken for conversions at the client end. This allows the server to make a balanced decision to send a particular format when several are available, minimising the information degradation and extra delay associated with format conversion.

The returned document contains an HTML header, and a body which may be in any notation or combination of encoding schemes which the client has declared itself able to handle. Caching of converted documents at client or server side is obviously a useful technique which could in principle be applied as an optimization.

## Data encoding

SGML was chosen as the base format for the returned document because it is an accepted encoding scheme and has traditional use in documentation. It is flexible enough to allow different object types to be encoded, and non-SGML encodings to be encapsulated. SGML was not chosen out of any particular aesthetic appeal or inherent cleanliness.

In general the philosophy of the W3 initiative is not to force any more standards or common practices upon the world than absolutely necessary. The document naming and addressing scheme is considered essential, as its flexibility allows us to include new protocols and name spaces as they appear. The protocols are not therefore seen as so essential, allowing continuing research and development. Transition strategies (using gateways for example) will allow the introduction of new protocols, and in a "market" of several coexisting protocols, the technically superior ones will hopefully achieve the most widespread use.

The data format negotiation of HTTP is designed to allow the same market forces to exist in data encoding schemes and document types. When two coworkers in the same field or organization are using the same tools, one imagines that the negotiation will allow them to exchange data at a high level, with a high level of functionality. For example, genetics workers may be able to exchange coded forms of DNA sequences which may be viewed and operated on with special tools. On the other hand, a worker in a very different environment will still have access to this data, even if as a last resort it has to be rendered into a plain ASCII text on his terminal. It will, after a while, become obvious which formats are becoming popular, and a snowball effect should cause their rapid adoption by both servers and clients.

The W3 clients will not themselves be able to handle all formats, but will be configurable to launch other applications.

All W3 clients are as a minimum required to handle plain text and HTML. Because HTML is a high-level mark-up, it allows the same logical structured text to be represented optimally whatever the capabilities of the client platform. For example, to highlight headings, a dumb terminal browser may use capital letters when an X-windows browser uses a different font.

## W3 software

Software provided by the various contributors to the W3 project includes browsers, servers and gateways.

A prototype hypertext editor was made using the NeXTStep(TM) environment. This has been frozen for almost a year at the time of writing, so it lacks many features of other browsers, but is still the only hypertext editor allowing links and annotations to be directly inserted by the reader or author.

The simple line mode browser "www" originally written by Nicola Pellow has now become a general information access tool. As well as operating in interactive mode, it can be used from the command line to retrieve any object on the web by name, or query any index. It can return formatted text or HTML source. When used as a filter, it becomes a text formatter. When used as a server, it becomes a gateway.

For the X11 window system, two browsers exist, with different look and feel. The "Erwise" browser was developed by a team at the Helsinki Technical University, and the "Viola" browser (demonstrated at this conference) by Pei Wei of Berkeley, California.

A browser for the Macintosh environment is under development at CERN, and an MSDOS/Windows browser may also be produced. As the status of this software is constantly being updated, it is wise to check the web for the latest situation.

On the server side, a simple file server exists. There are also examples of servers which have been set up using simple shell or perl scripts (see [2]). Contributions of new software are always welcome, as are new servers.

## Making data available

If you look at existing information in or about your own institute, or a particular field, maybe it occurs to you that it would be useful to others to make it generally available. There are a number of ways to do this, and it can be done in a few hours or a few days depending on the complexity of the data you have and the facilities you want to offer.

If you have an existing database, you could make a simple script to make this available on the web. The simplest way is to run an existing shell interface to the database from your server script. A more sophisticated way is to take that user interface program, and our skeleton server code, and merge them into a single C program to provide the data. This may be more powerful, or match better your style of working. A solution we used under VM/XA was for a basic daemon program to call a command file (REXX exec).

If you have no database, but some files of interesting information, then running the basic HTTP file daemon (httpd) will allow them to be made visible. In this case you can make by hand (or generate if you really have no time) a hypertext list of files for users to browse through.The format of HTML is described in the web. You can also pick up the source of any document you

find on the web with the -source option to the www (line mode browser) command.

## Research directions

The W3 project is not a research project, but a practical plan to implement a global information system. However, the existence of the web opens up many interesting research possibilities. Among these are new human interface techniques for managing a large space and the user's view of it, and automatic tools for traversing the web and searching indexes in pursuit of the answers to specific questions.

## More information

All the technical information about the W3 initiative is on the web. You can read it by telnet to info.cern.ch which gives you the simplest form of browser. Better, pick up a browser by anonymous FTP to the same host (directory /pub/www/src) and run it on your own machine. The www team at CERN will be happy to answer questions and receive feedback, ideas, requests and suggestions by email at: www-bug@info.cern.ch.

## References

[1] T.J. Berners-Lee et al., World-wide web: the information universe, in: *Electronic Networking: Research, Applications and Policy*, Vol. 2, No. 1 (Meckler, New York, 1992) 52–58.

[2] T.J. Berners-Lee et al., World-wide web: an information infrastructure for high-energy physics, in: D. Perret-Gallix, ed., *Proc. International Workshop on Software Engineering and Artificial Intelligence for High Energy Physics*, La Londe, France, January 1992.

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

[4] M. McCahil et al., The Internet Gopher: An Information Sheet, in: *Electronic Networking: Research, Applications and Policy*, Vol. 2, No. 1 (Meckler, New York, 1992) 67–71.

[5] B. Kantor and P. Lapsley, A Proposed Standard for the Stream-based Transmission of News, Internet RFC-977, 1986.

[6] B. Kahle et al., Wide Area Information Servers: An Executive Information System for Unstructured Files, in: *Electronic Networking: Research, Applications and Policy*, Vol. 2, No. 1 (Meckler, New York, 1992) 59–68.