World-wide web: the information universe

Tim Berners-Lee, Robert Cailliau, Jean-François Groff and Bernd Pollermann
CERN – European Organization for Nuclear Research, Geneva, Switzerland

Abstract

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

Design/methodology/approach – Since Vannevar Bush's article, men have dreamed of extending their intellect by making their collective knowledge available to each individual by using machines. Computers provide us two practical techniques for human-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. The W^3 ideal world allows both operations and provides access from any browsing platform.

Findings – 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 since it includes all anonymous FTP archives, WAIS servers, and Gopher servers as well as specific W^3 servers.

Originality/value – The paper notices that a W^3 server could provide the functions of each of these servers, and so it looks forward to a single protocol that can be used by the whole community.

Keywords Worldwide web, Internet, Communication technologies

Paper type Conceptual paper

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 then 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 (1945), 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 human-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

This article has been republished as part of *Internet Research's* commemorative 20th anniversary issue. It was originally published in *Electronic Networking* (later renamed *Internet Research*) Vol. 2 No. 1, 1992.

Internet Research
Vol. 20 No. 4, 2010
pp. 461-471

1066-2243
DOI 10.1108/10662241011059471

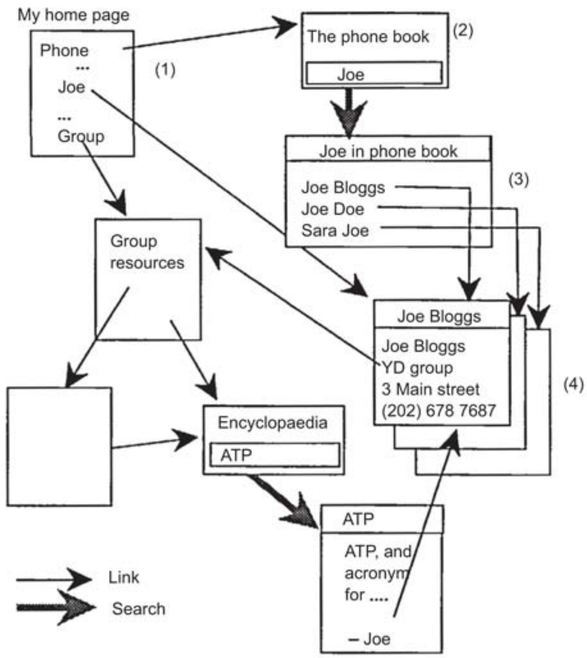
some words representing that which he desires. The W^3 ideal world allows both operations and provides access from any browsing platform.

Reality

Existing research projects and commercial products are not far form achieving parts of this dream. The Xanadu system is an ambitious distributed hypertext project. Existing hypertext systems (see for example *Beyond Hypertext, 1990*; Kahn and Meyrowitz, 1988; Nelson, 1990) tend to be restricted to the local or distributed file system and they often are developed with a limited set of platforms in mind. Contemporary information retrieval and access systems such as Alex (Cate, 1992), Gopher (Alberti *et al.*, 1991), Prospero (Neuman, 1992), and WAIS (Davis *et al.*, 1990) cover a wide area without the hypertext functionality. Merging the techniques of hypertext, information retrieval, and wide area networking produces the W^3 model.

The W^3 data model

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



Notes: The W^3 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 possiblity of inputting search words. The result is a virtual hypertext document (3) which points to the documents found (4)

Figure 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 is 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" that 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 snap-shots of changing data (such as the weather forecasts, 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 that 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 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; and
- common data formats for hypertext.

Most research in hypertext systems (the Xanadu project excepted) have focused on the user interface and authoring questions rather than on 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) that share a common file system. However, the W^3 architecture must cope with a widely distributed heterogeneous set of computers running different applications that 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 (Figure 2).

It may be more difficult initially to develop a generic hypertext browser than a specific front-end for a particular information system. However, the decoupling 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 from 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) that are not yet hypertext-aware. Therefore, ideally it must 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 and some "publishing" organization rather than any physical location, so that its location may later be moved.

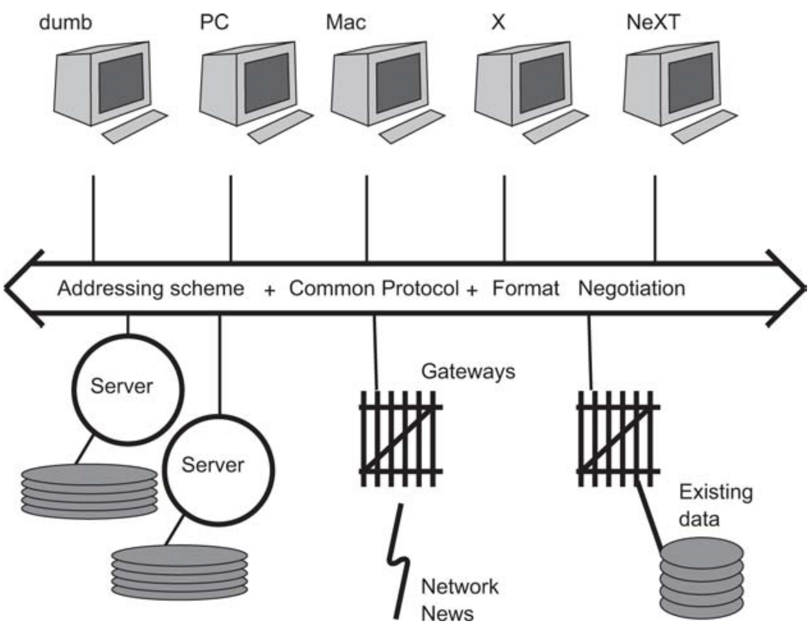


Figure 2.
The W^3 architecture in outline

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 that is transitory, such as the particular formats available for a document or its length.

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 who 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 (Postel and Reynolds, 1985).
- Network News Transfer Protocol – NNTP (Kantor and Lapsley, 1986).
- Access to mounted file systems.

A new search and retrieve (SR) protocol, known as HTTP, was found to be necessary. Faster than FTP for document retrieval, HTTP 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 (Halasz and Schwartz, 1990) provides a conceptual model for hypertext systems and the HyTime standard (Goldfarb, 1991) formalizes hypertext at a high level. The W^3 project defines a concrete syntax in the SGML style for basic hypertext as it is used for menus, search results, and online hypertext documentation.

Even W^3 browsing application is able to parse this simple format (see Figure 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 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 W3 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 also uses separate "source" files to describe indexes. The WAIS-W3 gateway keeps caches of these files, using them to build descriptive "cover pages" for indexes.

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 of this approach is that, as Yeong (1991a) points out, a large multimedia

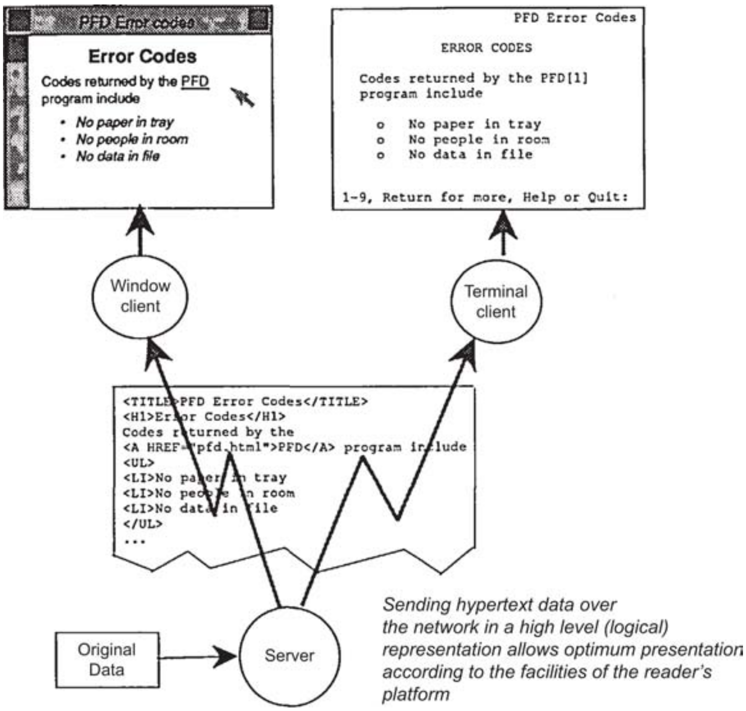


Figure 3.
A schematic illustration of the encoding of hypertext data. The link is represented in the window by underlining, on the terminal by a reference number

document may be most efficiently retrieved from a different host and by using a different protocol from that used for the original query. Furthermore, as online 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 a search, so that the next time he follows the link the search is re-evaluated. 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. First, users of-ten would like to be able to browse through available WAIS indexes. Both WAIS and W^3 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. Second, when one has found a piece of text, WAIS delivers just that part of file that has been found. Very often one would like links to surround 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 (Cate, 1992), Gopher (Alberti *et al.*, 1991), and Prospero (Neuman, 1992) 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, readers feel at home as they regard 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, because it uses 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 (1991b). 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 NeXT-Step™ 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 and 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 online, 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 hyper-text is 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 were important parts of the model. A typical information hunt will start from a default hyper-text 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 that 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 since it includes all anonymous FTP archives, WAIS servers, and Gopher servers as well as specific W^3 servers. We notice that a W^3 server could provide the functions of each of these servers, and so we look forward to a single protocol that can be used by the whole community.

The Archie project (Emtage and Deutsch, 1992) provides an index into the Internet archives and is an excellent example of a service that 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 W^3 access mechanism, 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 made possible 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 W3 protocols and data formats, and so forth, is available from our W3 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

Alberti, B., Anklesaria, F., Lindner, P., McCahill, M. and Torrey, D. (1991), "Notes on the Internet Gopher protocol", unpublished manuscript, Minneapolis: University of Minnesota, Micro-computer and Workstation Networks Research Center.

Beyond Hypertext: The DECWindows hyperenvironment vision (1990), Digital Equipment Corporation, Maynard, MA.

Bush, V. (1945), "As we may think", *The Atlantic Monthly*, Vol. 176 No. 1, pp. 101-8.

Cate, V. (1992), "Unpublished private communication", January 27, 1992, Carnegie-Mellon University.

Davis, F., Kahle, B., Morris, H., Salem, J., Shen, T., Wang, R., Sui, J. and Grinbaum, M. (1990), "WAIS interface Protocol: Protocol Functional specification", Thinking Machines Corporation, Park, CA, unpublished report.

Emtage, A. and Deutsch, D. (1992), “Archie – an electronic directory service for the Internet”, in Proceedings of the Winter 1992 USENIX Conference, USENIX Association, Berkeley, CA, pp. 92-110.

Gillies, J. and Cailliau, R. (2000), *How the Web was Born: The Story of the World Wide Web*, Oxford University Press, New York, NY.

Goldfarb, C.F. (1991), “Information Technology-Hypermedia/Time-based structuring language (Hy-Time) (ISO/IEC CD 10744)”, unpublished manuscript [available from authors].

Halasz, F. and Schwartz, M. (1990), “The Dexter Hypertext reference model”, in Moline, J. and Baronas, J. (Eds), *Proceedings of the Hypertext Standardization Workshop January 16-18, National Institute of Standards and technology, Gaithersburg, MD*, pp. 95-133.

Kahn, P. and Meyrowitz, N. (1988), *Guide, Hyper-Card, and Intermedia: A comparison of Hypertext/Hypermedia systems*, Brown University, Providence, RI, IRIS Technical Report 88-7.

Kantor, B. and Lapsley, P. (1986), “A proposed standard for the stream-based transmission of news (Internet RFC-977)”, February.

Nelson, T.H. (1990), *Literary Machines version 90.1.*, Mindfull Press, Sausolito, CA.

Neuman, C.B. (1992), “Prospero: a tool for organizing Internet resources”, *Electronic Networking: Research, Applications, and Policy*, Vol. 2 No. 1, pp. 31-8.

Postel, J. and Reynolds, J. (1985), File transfer protocol (FTP), Internet RFC-959.

Yeong, W. (1991a), “Towards networked information retrieval, (Technical report 91-06-25-01)”, Performance Systems International, Inc., Reston, VA.

Yeong, W. (1991b), *Representing Public Archives in the Directory*, Performance Systems International, Inc, Reston, VA, Internet draft.

Commentary on “World wide web: the information universe”

It is now more than 20 years ago since two people in the same research institute independently put forward a proposal for a networked hypertext. It was early 1989, one of the people was myself, the other was Tim Berners-Lee, whose proposal was more advanced and based on the internet protocols for carrying it. I joined Tim in 1990 and we worked closely together on what is now known as WWW.

As with most major technologies there was a lot of work going on elsewhere. One of these systems would ultimately have produced something very much like the web. But our project had two interesting qualities. It was the simplest of hypertexts since it had only the one-way link to connect documents. It also did not attempt to keep links working but accepted that over a set of independent servers there is no way to achieve coherence. Both of those qualities meant the web was scalable: it could grow without limits.

The article reprinted here is among the very early ones we wrote about the web, but we had little time to spend on academic explanations. There were two important pushes needed: one technological, mostly consisting of design and implementation of software and standards, and one social, mostly consisting of convincing people that the web was worth doing and that it would succeed only through collaboration. Being at least a decade older than the programmers around Tim, I occupied myself mostly with this second push.

Communicating the importance of a new idea is extremely difficult, especially to non-technical minds. But even among the geeks it was very much like the six blind men who try to describe an elephant:

Each one of these young and eager programmers had a different and incompatible technical goal. For the managers at CERN and elsewhere, the web might as well have come from outer space. I remember a meeting at the European Commission, where a very intelligent project manager still could not grasp the idea after a long explanation.

Words were no good, but to give a demonstration I had to arrange a meeting at the nearby Free University of Brussels: at that time only universities were connected to the internet.

There were lots of missed opportunities, which I will not recount, most are described in the book *How the Web was Born* (Gillies and Cailliau, 2000). Companies missed out on chances to work with the CERN team, we ourselves did not understand some basic social aspects, and even as late as spring 1995 Microsoft (though perhaps not Bill Gates himself) thought it was just another fad that would soon fade away.

There were disagreements inside the early development team, one was about the inclusion of a programming language. We know the result of leaving that hole in the design: it got filled by the world's worst programming language, Javascript.

Perhaps the biggest mistake was to distribute the so-called Line Mode Browser (LMB) in 1991. The web software Tim had implemented on the NeXT systems was brilliant in an some aspects still unsurpassed, but it could not be ported easily to the then in comparison archaic systems like Windows, MacOS and the various Unixes with their awkward or non-existent user interfaces. We hired a student to write a seriously downgraded program that could only browse, did not require anything beyond a 80x24 character grid display, not even cursor control, let alone a mouse. The LMB was written in minimal C. It could be downloaded and adapted to almost any platform in less than an afternoon. But it showed a very rudimentary web, which led lots of developers to dream up "improvements" that we actually already had in the NeXT version. It would be interesting to calculate, if it could be done, how much effort was wasted by the resulting chaos and how many years of deployment time were lost.

One crucial event, in which I played an important role, was CERN's decision to place the web technology in the public domain by an official document, signed on 30 April 1993. Without that, we would probably have seen attempts to fragment the web into proprietary (and therefore incompatible) islands. But that document, as important as it was, did not suffice in itself. During the years 1994 to 1997 much effort was put into the building of the Web Consortium. This could not be done without Tim and without serious support from academic and government sources. The collaboration, at times uneasy, between CERN, the European Commission, MIT and INRIA, did have the required effect of focusing the attention of all important players on the main issue: safeguarding the interoperability of all browsers and servers through the standards of http, html and all the technologies that followed.

But the web is based on the Internet, an academic development originally used by a very homogenous social group: academics. It has basic flaws when it is let loose on broader society. There is no identity, no tracing, no authentication, no payment system. The net flows freely through legal system boundaries, it is worldwide.

There is no way of enforcing local civil or legal behavior. The net and the web need legal agreements that are equally world-wide before we can solve those problems, because they are societal problems that no technology short of physical disconnection can tackle.

The use of www has now passed to social networking and the spreading of everyday information. The absence of a usable micropayment system with digital cash is the biggest problem I can see today. There is a vicious triangle of author-reader-advertiser. The authors have only a few ways of getting paid for their work. They can do it for free, but that is only an option for part-time activity. They can ask for a subscription fee paid up-front, with user names and passwords attached and perhaps some identification. They can use a small-payment system such as we see on music and application download sites. Finally they can try to get money from advertisers instead of the actual consumers. Advertising revenue is unstable and also often influences what information is offered. A completely automatic system of digital cash with micropayments would make a real market of information possible. It would essentially work much like the mobile phone system: you do not need to think about how your call is going to be paid for, there is a mature system behind it that will send you a bill at the end of the month. So, we are already used to such a mode of payment (indeed, for the information market it was implemented on the French Minitel system since the early 1980s). The entire web landscape will change for the better once the consumer can pay the author directly for quality information delivered with no advertising strings attached.

Finally a word on social networking: the most popular ones are quite frightening. You can get in, but you cannot get out. This is much like some religious groups, which you can join, but they will come to kill you if you dare become an apostate. Most people have no clue where their data are kept, where the companies involved are located, let alone which legal system would apply. They would not trust a used car salesman, but social networking sites are far worse.

Curiously, people worry about the bad uses the government could make of databases of personal details, but they seem not to worry about what businesses could do, or are already doing. Privacy laws are different in each country, and usually not enforced, so we do not know what is happening. On the internet, you do not know what is hidden from you.

Putting one's private documents in the "cloud" is touted as the way to go, by people who would not be seen dead in public transport.

There are huge problems brewing under the calm surface.

Humankind is in for a big surprise, real soon now.

Robert Cailliau

Formerly of CERN – European Organisation, for Nuclear Research, Geneva, Switzerland

About the author

Robert Cailliau graduated from Ghent University (Belgium) with a degree in Mechanical and Electrical Engineering, and from University of Michigan (USA) with an MSc in Computer, Information and Controls Engineering. From 1974 he worked at CERN, the European Laboratory for Particle Physics, where he was active in control systems, office computing, document handling and public communication of science. He spent the years 1989 to 1995 intimately involved in the creation of the Web together with Tim Berners-Lee. He was instrumental in making CERN put the web technologies into the public domain (1993), started the web technology conferences (1994), the Web for Schools project of the European Commission (1995) and helped transfer the web from CERN to the Web Consortium (1995-1996). He retired from CERN in 2007. He was awarded several international and national prizes for his role in the development of the web. Robert Cailliau is the corresponding author and can be contacted at: robert@robertcailliau.eu