

TimBL

## Commemorative Lecture

# The World Wide Web - Past Present and Future

Exploring Universality

### Abstract

The most important thing about the World Wide Web is that it is universal. By exploring this idea along its many axes we find a framework for considering its history, its role today, and guidance for future developments.

Hardware independence, which once meant running on mainframes, minicomputers and microcomputers, now extends to a multitude of devices from watches and speech devices to big screen televisions. The separation of the essential meaning of the information from the form in which it is conveyed helps this independence, and also makes the Web accessible to people with disabilities. Software independence, which is so important to prevent fragmentation into many disconnected proprietary webs, is under as much pressure as ever. That the Web must be independent of nation and location is nowhere so clear as in an international gathering such as this, where character set, language, and culture can be barriers which the technology helps us to bridge. As we look forward, we are tempted to distinguish between the multimedia world of information targeted for human perception, and the well-defined world of data which machines handle. A Web which encompasses both these extremes and all the rich land in between is the one which will help us best fulfill our hopes for society, for understanding between peoples, and finding a balance between the diversity and commonality in this rich world.

### Introduction

The concept of the Web integrated many disparate information systems, by forming an abstract imaginary space in which the differences between them did not exist. The Web had to include all information of any sort on any system. The only common idea needed to tie it all together was the *Universal Resource Identifier*(URI) identifying a document. From that cascaded a series of designs of protocols (such as HTTP) and data formats (such as HTML) which allowed computers to exchange information, mapping their own local formats into standards which provided global interoperability.

Back in 1989, before the World Wide Web, many different information systems existed. They ran on different sorts of computers, each running different operating systems, connected by different networks, and using quite different programs to give to the user very different ways of accessing information. Thus, while the information on two systems might be very relevant, the path between them was very long. And yet, in fact, each of the computer systems was very likely to be connected to some sort of network. And that network was very likely to be connected to another network, so that in fact there was a path from bit of data on one computer through a series of networks to the other computer. So there was, finally, no fundamental reason why these barriers to communication should exist.

The first breakthrough was the Internet, and I can't emphasize too often that I didn't invent the Internet! There were many networks, but they were of different types, some small, some large, and they used different sorts of connection. A computer could be on more than one network, and it was Vint Cerf and his colleagues who realized that a computer connected to more than one network could act as a kind of postal sorting office, and be used to forward information between the networks. Even though the little networks might use different numbering schemes for different computers, they imagined that each computer was on some global "Inter-network" and gave each computer a number. To describe things simply, the information is passed around in little packets (rather, as Vint says, like postcards) and each has on it a number that is the address of the computer to which it has to be delivered. The forwarding computers just look at the address number on each packet to figure out which network to send it over next. In this way, all you have to do is send off a packet with the right address number on it, and sooner or later it will arrive at the right place. The Internet was invented around the 1970s. I was fortunate in that in 1989, when I was looking at the problems of networked information systems, it was deployed across the US and to a certain extent in Europe.

The way the Web works is very simple. When you see a link in a Web page, it might be underlined, or blue, but however the computer indicates there is a link, that means that, in a special hidden code inside the document, there is the URI of the document to which the link goes. What happens when you click? Your computer looks at that URI, and if (like most URIs at the moment) it starts "http:", then it looks at the next bit, something like "www.w3.org". That is the *domain name* of the publishing authority, but what it needs is the number of a computer. Fortunately, a large number of *domain name servers* exist, computers which collaborate to hold a list of which domain name correspond to which computer address. Your Web browser sends, to one of these domain name servers, a packet containing the name of "www.w3.org", and received in return the address, "18.23.7.7". Your browser then sends off a packet to that address asking for the URI. The server responds by transmitting the new Web page back across the Internet to the browser. The browser receives the document, decodes the HTML tags in it, and displays a fresh Web page on your screen.

The Web required everyone to give a URI to their documents: a large request. To attain its universality, the design of the Web could not impose any extra constraint on how data was represented or organized. In fact, the first Web-specific communications protocol (HTTP) and

data format (HTML), designed at the same time as the URI, were very successful and used for a very large amount of the web. However, the Web still was designed to only fundamentally rely on one specification: the Universal Resource Identifier.

## Device independence

That the same information should be accessible from many devices is a core rule of the Web. Once the choices were 80 character terminals or the new personal computers. Now, the number of pixels on a computer screen has steadily increased, but mobile devices have small screens or voice input and output. Our ability to represent information independently of the hardware we use is more important than ever.

The direct impact of the Web was seen in its ability to cross hardware and software boundaries. Before the web, at CERN, academic papers and administrative data were kept on a mainframe computer, but much live information and "help" information was available on minicomputers. Most people had replaced the terminals in their offices with personal computers, but still kept a window open logged onto the mainframe simply to access the phone book. Unexciting though it was, access to the phone book from computers of all sorts was an early incentive for browser adoption at CERN: for these people it was the critical application which convinced them.

A crucial factor in the design of the Web was the use of markup languages which transmitted the intent of the markup instead of the actual form for display. For example, tags for *heading level one* rather than *centered bold big text* allowed the same information to be displayed on color terminals with only one font, as well as black and white multifont windows, or whatever was available. This concept, the **separation of form and content**, is very important still for today's Web designers.

Interestingly, the Web spread so fast that it was not apparent to many designers how limiting it would be to make assumptions about which device a user had. Many sites proclaimed that they were "best viewed using 800x600 pixel screens". A few years later, as typical screens increased to 1024x768 pixels but many users were still using old 640x480 screens, the mistake became apparent.

More recently, the need for device independence has taken on new dimensions as the long-promised dawn of practical speech recognition software becomes a reality. Speech interaction breaks the user interface metaphor assumption which graphic user interfaces introduced, the idea that the computer and human share a view of a document. Speech interfaces bring us back to the conversational style which in fact computers used in the old days of the command line program. This change is more than just one of screen size. When we try to generalize a user's interaction in a way that may include mixtures all these modes, it causes significant

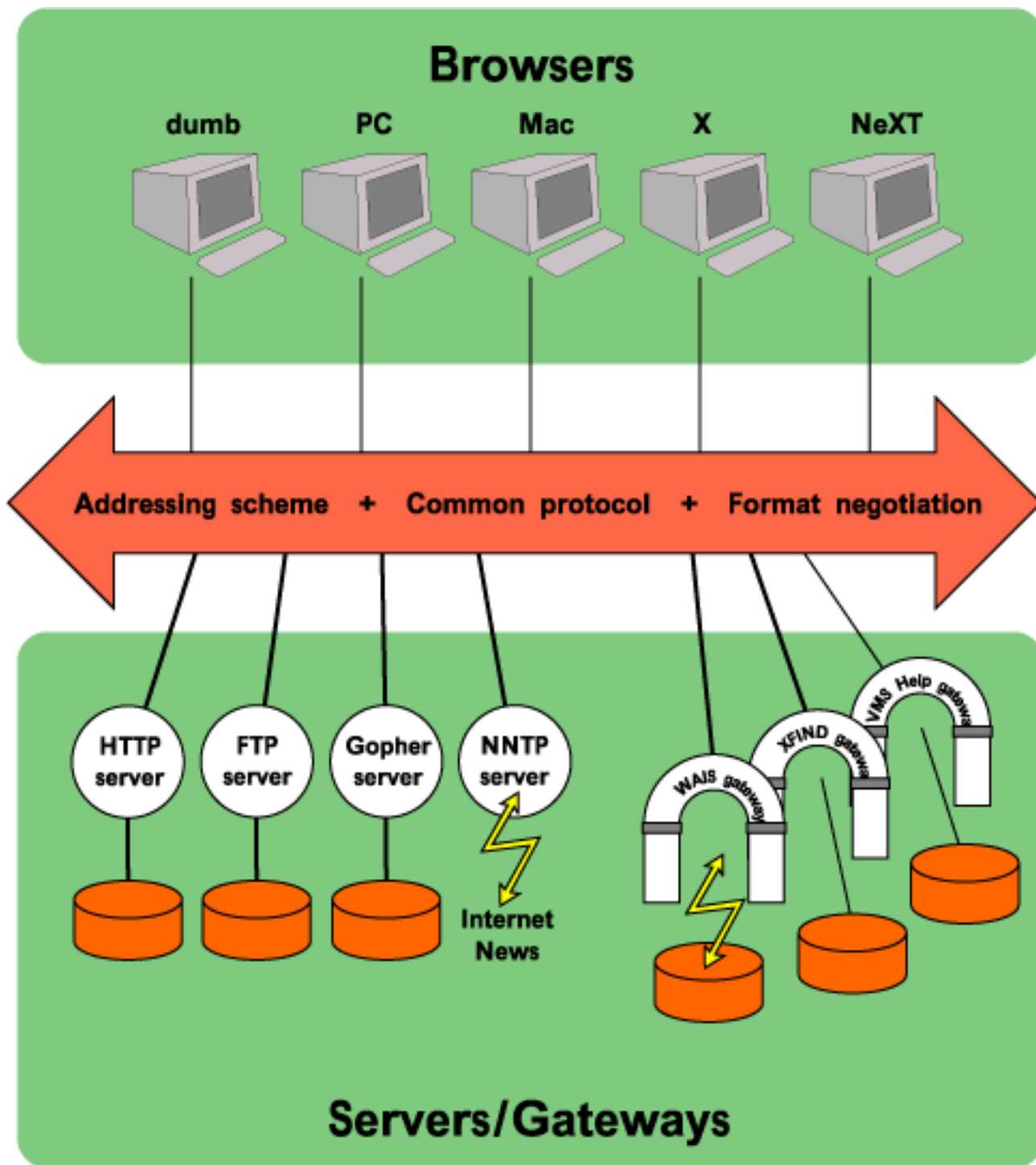
rethinking, in which the community is currently (2002) engaged.

## Software Independence

Many different forms of software provide and consume Web information, and no one program was critical to the whole Web. This decentralization of software development was and always will be crucial to its unimpeded growth. It also prevents the Web itself from coming under the control of a given company or government through control of the software. Communication standards give people a choice of software, but we must all learn to be aware of when our experience is being controlled by software with a bias.

The Web was deployed not as a program, but as a set of protocols.

The initial diagrams made it clear that those specifications -- URI, HTTP, HTML, and others -- would form a sort of "bus" connecting the many different sort of user programs ("clients") and many different sorts of information provider programs ("servers").



The initial client for the Web ran on a NeXT computer, at that time the most sophisticated platform available. The second client was a simple terminal-oriented command-line program for use on systems which didn't have a graphic interface at all. Between the two they demonstrated the concept of software independence.

The market situation around Web software has been through many phases, but this issue has always been important, and still is today. Now that so much money, and human attention - which is quickly turned into money, flows through the web-human connection, control of any aspect of the interaction with a human can be very lucrative.

Soon companies tried to find ways to influence and control the user's choice of information. Computers came with free software, and software came with built-in bias toward certain Web pages, and certain search engines. Users thinking they are just "searching the Web" use a specific search engine which points them to specific information, views, and products. Not only must the technology support a choice of software, but a competitive market must exist, and users must be informed and aware of what is going on.

## Internationalization

From its beginning in a laboratory run by over a dozen collaborating countries, the Web had to be independent of any inherent bias toward one given country. XML, being firmly based on Unicode, now allows all kinds of characters. Internationalization must take into account much more: the direction in which text moves across the page, hyphenation conventions, and even cultural assumptions about the way people work and address each other, and the forms of organization they make.

In 1994, in response to pressure for a body to coordinate development of interoperable standards for the Web became intense. The World Wide Web Consortium (W3C) was founded at the Laboratory for Computer Science (LCS) at the Massachusetts Institute of Technology (MIT) to lead the technical evolution of the Web and ensure its interoperability by developing common protocols. A lot of effort is spent at the W3C to justify the first two W's of "WWW".

The first HTML documents were unfortunately (due to my ignorance of *Unicode*) capable of representing only Western European languages. Since then, the new version, XHTML is based on XML, which is based on the *Unicode* standard.

Nowadays, the Consortium's Internationalization Working Group reviews new technology to try to spot areas in which a national, linguistic, or cultural bias may have crept into the design. We are very pleased to be hosted in parallel by INRIA in France and Keio University in Japan, as well as the Massachusetts Institute for Technology in the United States of America.

For all this work, the English language still tends to dominate the Internet. From the technical point of view, the Internet had been installed across the US when the Web started, but had not spread so much in other countries. From the market point of view, the US provides a single-language block which is a huge market for a new Web site, in contrast to Europe where the need for translation into many languages is a hinderance to the explosive uptake of a new site. I certainly hope that the Web will allow many cultures and languages to flourish, and that we will not sink to that common subset of expression which we can all understand.

## Multimedia

Multimedia is not just a buzz-word, it stands for an important dimension of variety - the palette of technologies available to human creativity. Even the early demos of the web included sounds and music. What has changed since then is that the capacity of typical computers to handle graphics and sound has increased, and for some, the bandwidth even allows video to be sent. Because many things can still be done with plain text, the exotic and the mundane will always coexist on the Web.

The first Web pages were displayed in a variety of fonts and formatting options, but images, sounds and movies were separate documents linked from the text page. Marc Andreessen's *Mosaic* browser led the way in integrating images, and Pei Wei's *Viola* browser demonstrated the power of scripting. Now, single Web pages can integrate text, photographs, line drawings, and mathematical formulae. The image technology has advanced with standards in *Scalable Vector Graphics* which allows a drawing to be sent as an abstract collection of graphic objects, and rendered on arrival into the appropriate style and resolution for each device, whether a large computer or a small phone. This gives much better results than the use of pixel graphics such as GIF and the later PNG. With the Synchronized Multimedia Integration Language (SMIL) there is now a standard for how all manner of multimedia things should be combined into a single experience. Unfortunately, in streaming audio, standards are less clear.

It is still the case, as a decade ago, that bandwidth and processor power limit what is practical, especially for video. But always, plain text, which needs neither of these, is all one needs for poetry and for most electronic commerce.

## Accessibility

Just as people differ in the language, characters and cultures to which they are used, so they differ in terms of their capacities, for example, in vision, hearing, motor or cognition. The universality which we expect of the Web includes making sure that, as much as we can, we make the Web a place which people can use irrespective of disabilities. There are guidelines for Web site designers to help with this now, and a site which follows them will typically be easier for anyone to use, and easier to index and search.

The separation of form and content, referred to above, is also a key to making the Web accessible to those with disabilities. To communicate well, we need not only to master each multimedia genre as effectively as we can, but we must also allow people a choice of medium, as users get on more easily with some than others. Soundtracks have subtitles, images have descriptions, mouse movements have keyboard alternatives, and so on.

It turns out that this work overlaps very much with other areas. Accessibility is enhanced when we have separated form and content, and the forms of available include different media. It

takes a bit of extra work, for example in using the text of a to the video in the form of captions for the visually impaired, and in making up textual explanations of the contents of images but it is important and worth the extra effort.

## Rhyme and Reason

There is another axis along which information varies. At one end of the axis is the poem, at the other the database table. The poem, or for that matter the 15 second TV commercial, is designed to connect to a human brain using all its complex series of associations in clever and powerful ways which we can never fully analyze. The database is designed to be queried and processed by a machine. It has well-defined values of information regularly arranged in columns which, hopefully, has well-defined meaning. Databases can be joined and split, combined and repurposed. Human beings use different sides of the brain for dealing with these types of information. Most information on the Web now contains both elements. The Web technology must allow information intended for a human to be effectively presented, and also allow machine processable data to be conveyed. Only then can we start to use computers as tools again.

When I first was at CERN, the computing division was known as the DD, or *Data and Documents* division. That name was later deemed outmoded, and the usual phrases such as *Management Information Systems*, *Computing and Networking*, and *Information Technology* were used. However, the old name draws a useful distinction. One can think of documents as information items, multimedia possibly, for people. Data, on the other hand, is for machines; hard, well defined, the stuff of computation.

The elegance of the WWW browser as a computer application was that it almost completely hid its workings from the user. The user never saw the HTML and, in the first browser, never saw the URIs. The job of the machine is to keep a low profile, to leave the user alone in an abstract space of documents. And so it should be, as machines cannot really do much else in the realm of documents. They cannot understand them, and therefore cannot work with them.

In the realm of data, things are different. Numbers can be crunched. Rules can be applied. Data can be sifted and correlated by machines very effectively. This is what the late Michael Dertouzos, Director of the LCS at MIT, described as the "heavy lifting" of information work. The analogy is with building work, like when machines can shift the earth much better than we can, though they add no creativity as to how to do it. The Web at the moment lets us down in the area of data, because the data is not in a form which machines can use. It isn't well identified in terms of the way it should be combined. All a computer can do is to pretend to be a person browsing the Web, and then guess what each Web page means!

The Semantic Web development adds to the Web formats for representing data and its

*semantics* - the meaning for a machine in terms of what rules can be applied and how it can be transformed into other data. This will lead to much greater clarity in complex communications, when an invoice is sent with some accompanying simple mathematics which describes its role in commerce transaction. It will lead to much greater re-use of data, and much easier analysis of what is going on.

## Quality

Many documentation systems used to be designed for particular collections of information, and one could assume that the information in such a system had achieved a certain quality. However, the Web itself cannot enforce any single notion of quality. Such notions are very subjective, and change with time. To support this -- to allow users to actually use the web even though it contains junk as well as gems -- the technology must allow powerful filtering tools which, combining opinions and information about information from many sources, are completely under the control of the user.

It is understood that a collection of works, such as a set of technical reports or a library, only includes articles reaching a certain standard, and some early dial-up information services similarly amassed information according to some quality criterion. Some people miss that with the Web -- hence the need for portals which provide a filtered view. However useful people find such portals, though, it is important that the Web itself doesn't try to promote a single notion of quality.

The Web has to be able to carry, uncomplaining, beauty and ugliness, honesty and lie. Users who find all of this of course complain, and sometimes ask for it all to be organized and filtered. However, not only would one central authority for quality be socially a disaster, but also, any one single categorization of data would be only one person's view. Human knowledge is not a tree, it is a web. How can we give the user the subjective perception of higher quality, while maintaining an open Web for people whose criteria are different?

The answer is through filtering. Unlike censorship, which is the forceful prevention of one person's communication by another, filtering is the control by the reader of what he or she reads. The trick is to allow the user to choose another person, or another group's, criteria of selection. This is what happens when a user selects from one of a choice of portals. More sophisticated systems involve white lists of "desirable" sites, or black lists of "undesirable" sites to be selected. This sort of information about information is known as metadata. Metadata in general includes all the information which catalogers and publishers and librarians keep about information. The Semantic Web languages (such as RDF) allow metadata to be exchanged freely between different parties. As the richness of metadata grows, so users will be able to combine criteria to hone their searches and guide their browsing. And the Web will be left unconstrained by a central authority deciding what information is appropriate for everyone.

There will, always be trash out there, and gems. Remember that you don't have to read the junk. And also remember that the unimportant notes of today maybe the foundation of revolutionary new ideas tomorrow.

## Independence of Scale

The Web is described as a global phenomenon, and it is, but we must remember that personal information systems, and family and group information systems are part of it too. There should be no information boundary which would prevent a link from my personal diary to a public meeting. We know we need harmony on a global scale for peace, but that peace will only be stable so long as social groups of all sizes are respected. Starting at the individual, a group of one, one can think of institutions and ad-hoc groups of all sizes. The Web must support all of those, allowing privacy of personal information to be negotiated, and groups to feel safe in controlling access to their spaces. Only in such a balanced environment can we develop a sufficiently complex a many-layered fractal structure which will respect the rights of every human being, and allow all the billions of us to live in peace.

When people express to me nervousness about the Web, there are two concerns I hear repeatedly.

The first is that the Web will become one giant MacDonalds, and international only as a mono-language, mono-cultural block. The French feared that the transatlantic Internet cable would cause the culture of the Louvre to be trampled over by the culture of Disney. People fear that only one portal will end up surpassing all the others and become the only lens through which all people see the world. It's a serious concern that if we have a global network, we will homogenise our culture. It would be horrible if language began to contain only those concepts that are sufficiently bland to be understandable by absolutely everybody. We would lose a great deal of richness. We need diverse pool of ideas for solving the unknown problems ahead of us as teh human race.

The other fear is the opposite. One can chose what Websites one reads. One can filter one's email, so that one sees no information except from a small group, a clique, or possibly a group of deluded dangerous fanatics. A person can operate in a virtual world without reality-checks from friends and neighbors. The danger for people who operate without interaction with the larger world is that the only common language they have with those different from themselves may be violence. Our world at the moment desperately needs enough common understanding to bring peace.

So, it is important as well, that while we have diversity, that there is a balance between the small scale culture and the large scale culture - and all scales in between. It seems that it is not only society which clearly needs a balance: in its way, a lot of nature does as well. Nature

is filled with fractal patterns. This can be seen for example in ferns or coastlines. One might approach a coastline, and as one gets closer above the coast, it has a certain interesting structure. Then, closer, to a tenth of the altitude, it still has an interesting structure. Closer and closer, until the point where the seaweed curling around a few of the pebbles is visible, it still has an interesting structure. It has structure in all levels. I have a deep feeling that society needs to be like that. It can't be a simple structure which operates just at one level. We need a complicated structure, which is fractal in some way. That means that our society and the technology which we use to support it has to work at each of these levels.

The development of the World Wide Web is a great example of human endeavor in which many people participated, driven by individual excitement and a common vision. There was no global management plan to make the World Wide Web. It happened because a very diverse group of people, connected by the Internet, wanted it to happen. The process was great fun, and still is. From the fact that it worked, I draw great hope for all our futures. May we now use every ability we have to communicate to build a society in which mutual respect, understanding and peace occur at all scales, between people and between nations.

---

Tim Berners-Lee

For Japan Prize Commemorative Lecture, 2002.

with thanks to Amy van der Heil for helping to put this together.

Last change \$Id: Lecture.html,v 1.9 2003/01/22 18:59:17 amy Exp \$

[TimBL](#)