II Weaving the World Wide Web

We should work toward a universal linked information system, in which generality and portability are more important than fancy graphics techniques and complex extra facilities. The aim would be to allow a place to be found for any information or reference which one felt was important, and a way of finding it afterwards. The result should be sufficiently attractive to use that the information contained would grow past a critical threshold.

Tim Berners-Lee[1]

The hypertext visionaries

Vannevar Bush (B.11.1), creator of the "Differential Analyzer" machine, wrote the very influential paper "As We May Think" in 1945, reflecting on the wartime explosion of scientific information and the increasing specialization of science into subdisciplines:

> There is a growing mountain of research. But there is increased evidence that we are being bogged down today as specialization extends. The investigator is staggered by the findings and conclusions of thousands of other workers – conclusions which he cannot find time to grasp, much less remember, as they appear. Yet specialization becomes increasingly necessary for progress, and the effort to bridge between disciplines is correspondingly superficial.[2]

Bush concluded that methods for scholarly communication had become "totally inadequate for their purpose."[3] He argued for the need to extend the powers of the mind, rather than just the powers of the body, and to provide some automated support to navigate the expanding world of information and to manage this information overload. For this purpose, he introduced the idea of a new type of device:

> Consider a future device for individual use, which is a sort of mechanized private file and library. It needs a name, and to coin one at random, "memex" will do. A memex is a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory.[4]

B.11.1. Vannevar Bush was born in 1890 in Massachusetts, attended Tufts College, and obtained his PhD in engineering from MIT in 1917. He joined the MIT Department of Electrical Engineering two years later and in 1927 started work on his Differential Analyzer – an analog computer for solving complicated systems of differential equations. One of his graduate students was Claude Shannon whose MIT master's thesis on using electrical relays to implement Boolean logic operations is one of the most-cited MIT theses. During World War I, Bush had been frustrated by the poor cooperation between the U.S. military and civilian scientists. When World War II broke out, he persuaded President Roosevelt to set up the National Defense Research Committee with him as chairman. Bush later said "if he made any important contribution to the war effort at all, it would be to get the Army and Navy to tell each other what they were doing."[B1] Two of the most celebrated technological successes that were overseen by Bush were the Manhattan Project – that produced the atomic bomb – and the "proximity fuse" – a fuse inside an artillery shell that contained a miniature radar system so that the shell would explode when near the target. After the war, Bush's report to President Truman – "Science: The Endless Frontier" – advocated federal funding of civilian basic research and resulted in the establishment of the National Science Foundation in 1950. He also wrote a second famous paper in 1945 that was titled "As We May Think." It is surprising how many of the visionary ideas described in this paper are still relevant in the era of the World Wide Web, almost seventy years later.

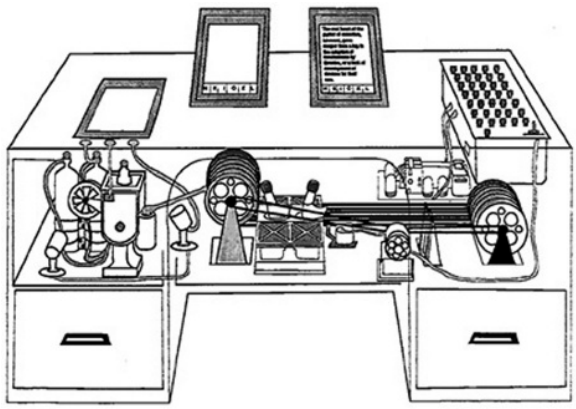


Fig. 11.1. An illustration of the memex – a personal information system envisaged by Vannevar Bush to help users cope with the increasing flood of information.

Although the specific technology proposed by Bush to create the memex is now hopelessly out of date (see Fig. 11.1), his idea of recording "links" to represent associations between two pieces of information was the inspiration for today's World Wide Web. By using such links, Bush thought we could mimic the working of the human mind in how it follows a trail of associations. This is how he imagined the memex machine working:

> The owner of the memex, let us say, is interested in the origin and properties of the bow and arrow. Specifically he is studying why the short Turkish bow was apparently superior to the English long bow in the skirmishes of the Crusades. He has dozens of possibly pertinent books and articles in his memex. First he runs through an encyclopedia, finds an interesting but sketchy article, leaves it projected. Next, in a history, he finds another pertinent item, and ties the two together. Thus he goes, building a trail of

Fig. 11.2. The cover of Nelson's 1974 book *Computer Lib*. The book was published together with another book – *Dream Machines* – which was about the media-handling potential of computers.

> many items. Occasionally he inserts a comment of his own, either linking it into the main trail or joining it by a side trail to a particular item. When it becomes evident that the elastic properties of available materials had a great deal to do with the bow, he branches off on a side trail which takes him through textbooks on elasticity and tables of physical constants. He inserts a page of longhand analysis of his own. Thus he builds a trail of his interest through the maze of materials available to him. And his trails do not fade.[5]

This was essentially the first description of *hypertext*, a text with interactive connections – *hyperlinks*, as we call them now – that give many options for moving between documents and allow a reader not to be restricted to just following a "linear" path through a document. It was this vision that inspired Doug Engelbart to start building his oN-Line System (NLS) in 1962. Engelbart's NLS was the earliest working hypertext system, complete with mouse, and he first demonstrated it in public at the famous "Mother of All Demos" in San Francisco in 1968.

Bush's ideas also inspired another visionary, Ted Nelson (B.11.2). Nelson was studying for a master's degree in sociology at Harvard University in 1960 when he attended a course on computers. It was then that he conceived of building a software system that would allow nonsequential editing and reading of documents and permit the composition of compound documents that would display pieces of other documents, a concept he called *transclusion*. In a lecture at Vassar College in February 1963 Nelson described a system he called PRIDE, for *personalized retrieval indexing and documentary evolution*, that would enable the user to organize many types of research and personal notes. It was in this lecture that Nelson introduced the term *hypertext* (initially spelled with a hyphen) to describe his ideas about how to overcome the linear constraints of ordinary documents. The essence of hypertext was that it enabled nonlinear writing and reading, so that the reader could jump to another location in the text or move to a completely different document (Fig. 11.2).

Nelson refined his ideas over the next few years into an ambitious project he called Xanadu. He envisaged this as a system that supported two-way links; had unique and secure identifiers for users and documents; and provided a mechanism for tracking "micropayments" to authors for the use of parts of their writings. Nelson hired programmer Cal Daniels to produce a demonstration system of Xanadu in 1972. Two years later Nelson updated his vision to include networked computers and a repository for information that he called

B.11.2. Ted Nelson is both a visionary and a pioneer of many innovative ideas. He was born in New York in 1937 and has degrees from Swarthmore College in philosophy, from Harvard in sociology, and from Keio University in media and governance. Nelson has made important contributions to computer science with his ideas about hypertext – text that is linked to other information – and about new types of documents. Nelson also strongly supported personal computing with the rallying cry "Down with Cybercrud," protesting the centralization of computers. In 1974, prior to the release of the Altair personal computer, he published the book *Computer Lib* with the subtitle *You Can and Must Understand Computers NOW*.

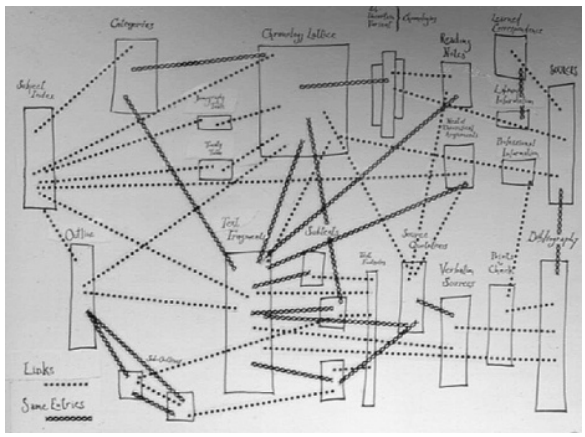


Fig. 11.3. Sketches of Ted Nelson's early global hypertext system from 1965. He also referred to this idea as a document universe or "docuverse." There are two types of links in this picture: the dotted lines represent normal hyperlinks, and the braided lines, representing links that point to quotations from other documents, Nelson called "transclusion" links. The system also introduced the idea of parallel text that enables us to see several related documents at the same time, as if we had several pages in front of us on a desk.

the *docuverse* (Fig. 11.3). However, it was not until 1998 that the first, still incomplete, Xanadu system was released and by then, the growth of the World Wide Web was already well under way. Although the present-day World Wide Web incorporates some aspects of his vision, Nelson calls Tim Berners-Lee's version of hypertext "precisely what we were trying to PREVENT – ever-breaking links, links going outward only, quotes you can't follow to their origins, no version management, no rights management."[6]

Vannevar Bush concluded his 1945 article with a surprisingly accurate vision of today's world of information. He accurately foresaw the emergence of such things as Wikipedia and social networks but totally missed the central role now occupied by Internet search engines:

> Wholly new forms of encyclopedias will appear, ready made with a mesh of associative trails running through them, ready to be dropped into the memex and there amplified. The lawyer has at his touch the associated opinions and decisions of his whole experience, and of the experience of his friends and authorities. The patent attorney has on call the millions of issued patents, with familiar trails to every point of his client's interest. The physician, puzzled by a patient's reactions, strikes the trail established in studying an earlier similar case, and runs rapidly through analogous case histories, with side references to the classics for the pertinent anatomy and histology. The chemist, struggling with the synthesis of an organic compound, has all the chemical literature before him in his laboratory, with trails following the analogies of compounds, and side trails to their physical and chemical behavior.
>
> The historian, with a vast chronological account of a people, parallels it with a skip trail which stops only on the salient items, and can follow at any time contemporary trails which lead him all over civilization at a particular epoch. There is a new procession of trail blazers, those who find delight in the task of establishing trails through the enormous mass of the common record. The inheritance of the master becomes, not only his additions to the world's record, but for his disciples the entire scaffolding by which they were erected.[7]

Fig. 11.4. Bricks-and-mortar libraries were the traditional way of doing research before the advent of the World Wide Web. This photograph shows the famous Reading Room at the British Library in London. People wanting to use it had to apply in writing for a reader's pass, which would be issued by the Principal Librarian.

In June 1992, in an issue of the University of Minnesota *Wilson Library Bulletin*, librarian Jean Armour Polly wrote an article titled "Surfing the Internet." In the article, she described how she could, from her home in New York, "surf" from server to server looking for information across the world:

> Today I'll travel to Minnesota, Texas, California, Cleveland, New Zealand, Sweden, and England. I'm not frantically packing, and I won't pick up any frequent flyer mileage. In fact, I'm sipping cocoa at my Macintosh. My trips will be electronic, using the computer on my desk, communications software, a modem, and a standard phone line.[8]

Polly got the idea for the metaphor from a picture of an "information surfer" on her Apple Macintosh mouse pad. At the time, "surfing the Internet" was a very labor-intensive process. It meant explicitly downloading files from remote servers using the *file transfer protocol* (FTP). What finally made information surfing easy was the combination of two developments. One was the World Wide Web developed by Tim Berners-Lee and Robert Cailliau at CERN, the high-energy physics research laboratory in Geneva, Switzerland. The other was the easy-to-use Mosaic browser, the first web browser to gain popularity among the general public. Mosaic was created by Marc Andreessen, then still a student, and Eric Bina, a staff member at the National Center for Supercomputing Applications (NCSA) at the University of Illinois. The Mosaic browser included features such as *icons* (tiny pictures), *bookmarks* to store locations users might want to revisit, and a simple point-and-click method for finding, viewing, and downloading information that was appealing to people with little knowledge of computers.

Librarians were among the first people to see the impact of the Internet on how we access knowledge. Until recently, bricks-and-mortar libraries served as the temples of knowledge (Fig. 11.4), but with the arrival of the Internet and the web this has changed. We can now literally have all the books and all the information in all major libraries in the world at our fingertips. The invention of the World Wide Web and the technologies used to search it are the subject of this chapter.

B.11.3. Tim Berners-Lee graduated from Queen's College, Oxford in 1976 with a degree in physics. In 1980, at the CERN Laboratory near Geneva, Switzerland, he started to work on ideas for a novel "web" of information. His work was motivated by the need to develop a system that would provide fast access to manuals describing complex equipment, experiments, and other documentation used at CERN. Berners-Lee's great contribution was to devise an engineering solution for combining the Internet and *hypertext* links into a powerful tool. *Time* magazine named him as one of the twenty most influential persons of the twentieth century and in 2003 he received a knighthood from the Queen.

Error 404 and the World Wide Web

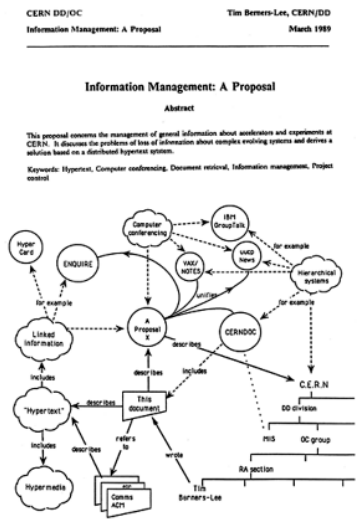
In 1980, Tim Berners-Lee (B.11.3), a young physicist turned software engineer, accepted a temporary software consulting position at CERN, the famous European laboratory for particle physics near Geneva, Switzerland (Fig. 11.5).

Fig. 11.5. An aerial view of the CERN laboratory on the France-Switzerland border just outside Geneva, Switzerland. The large circle shows the location of the tunnel for the Large Hadron Collider, the world's largest particle accelerator. This machine made possible the discovery of the Higgs boson by particle physicists in 2012.

CERN was then in the process of updating the control systems for its particle accelerators, devices that propel subatomic particles to high speeds, and Berners-Lee had been hired to help. For a temporary contract programmer, it was a major challenge to understand all the different components of the control system and to know who was responsible for each component. To help him keep track of all this information, Berners-Lee wrote a software program called *Enquire*. The name was a shortened version of the title of a how-to book called *Enquire within upon Everything*, a Victorian compendium of household advice that he remembered from his childhood. In his Enquire system, Berners-Lee could input a page of information about a person, a device, or a program, as he explained:

> Each page was a "node" in the program, a little like an index card. The only way to create a new node was to make a link from an old node. The links from and to a node would show up as a numbered list at the bottom of each page, much like the list of references at the end of an academic paper. The only way of finding information was browsing from the start page.[9]

The program stored data in a much easier to use way than a traditional rigid hierarchal organization by allowing links to different paths of information. It had two types of links, an "internal" link within a file and an "external" link that could jump between files. The external link went in only one direction, which was important to avoid the problem of many people linking to the destination page and the owner having to store many thousands of return links. This initial program did not run on a network of computers but on a stand-alone computer.


CERN DD/OC Tim Berners-Lee, CERN/DD
Information Management: A Proposal March 1989

Information Management: A Proposal

Abstract

This proposal concerns the management of general information about accelerators and experiments at CERN. It discusses the problems of loss of information about complex evolving systems and derives a solution based on a distributed hypertext system.

Keywords: Hypertext, Computer conferencing, Document retrieval, Information management, Project control

Fig. 11.6. Tim Berners-Lee's original 1989 proposal for a novel hypertext-based document management system at CERN. The figure shows circles with arrows, indicating documents, organizations, and people all connected by electronic links called hyperlinks.

Berners-Lee left CERN after about six months but returned to their Computing and Networks Division four years later. By this time, the researchers working on large particle physics experiments routinely networked their computers together, connecting the scientists working at CERN with one another and with their home institutions. After a year or so mulling over the new networked environment at CERN and his previous experiences with Enquire, Berners-Lee decided that a new type of "document management system" was needed that would essentially be a version of hypertext that operated over the Internet. In order that users would not be required to obtain access permissions from other users, the system would need to be completely decentralized with no center of control and with no one keeping track of all the available links. This decentralization was important because Berners-Lee was aware that this was the only way that the system could scale to accommodate thousands and even millions of users. Furthermore, he said, "the act of adding a new link had to be trivial"[10] – because the easy addition of links was vital if his "web" of links was to spread around the world. In late 1988, Berners-Lee talked to his boss, Mike Sendall, who encouraged him to write up a proposal that would establish his idea as a formal project (Fig. 11.6).

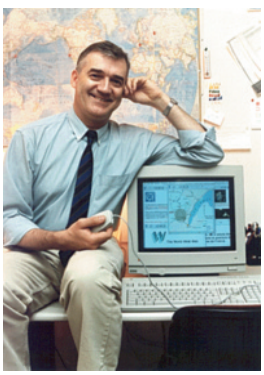
At that time, Vint Cerf and Bob Kahn's TCP/IP networking protocol for the Internet was not well established in Europe. Nevertheless Berners-Lee chose to adopt their protocol because the particle physics community used and loved the Unix systems that all supported TCP/IP for network communications. In March 1989, he gave a version of his proposal to Mike Sendall and Sendall's

boss, David Williams. Williams had the difficult job of defending his Data and Documentation Division – renamed in 1990 as Computing and Networks Division – from accusations by physicists that the division was wasting money on exotic computer science research that could better be spent supporting physics experiments (B.11.4). Berners-Lee's networked hypertext project was therefore only "informally" allowed to proceed. In 1990, with Williams deliberately turning a blind eye to the project, Sendall authorized the purchase for Berners-Lee of a new NeXT computer, a high-end workstation developed by Steve Jobs. Sendall told Berners-Lee, "When you get the machine, why not try programming your hypertext thing on it?"[11]

B.11.4. Official CERN photo of Peggie Rimmer, the manager in CERN's Documentation and Data Division who was responsible for hiring Tim Berners-Lee back to CERN in 1984. She, together with her husband, Mike Sendall, and the Documentation and Data Division director, David Williams, provided the necessary "air cover" from the physicists to allow Tim Berners-Lee to invent the web.

Berners-Lee was now faced with the uphill task of convincing the CERN scientists, who were extremely busy and already highly proficient in the use of computers, that his idea of a global hypertext system was both exciting and useful. First he needed a name. After debating possibilities such as "Information Mesh" and "Mine of Information," he settled on the ambitious-sounding "World Wide Web." He was then fortunate to find the ideal evangelist to help him spread the message to the CERN community, the Flemish-speaking Belgian Robert Cailliau (B.11.5). Berners-Lee said later, "In the marriage of hypertext and the Internet, Robert was best man."[12] They began by trying to interest some of the companies that currently sold nonnetworked hypertext systems. Even Electronic Book Technologies, Inc., in Rhode Island, established by the legendary computer scientist Andy van Dam from Brown University, who had earlier collaborated with Ted Nelson, rejected Berners-Lee's proposal. Van Dam instead insisted on a centralized link database so that there would be no broken links. By contrast, Berners-Lee envisioned a truly dynamic system:

> I was looking at a living world of hypertext, in which all the pages would be constantly changing. It was a huge philosophical gap. Letting go of that need for consistency was a crucial design step that would allow the Web to scale.[13]



B.11.5. Robert Cailliau was born in Tongeren, Belgium in 1947 and has degrees in electrical and mechanical engineering from the University of Ghent and in computing from the University of Michigan. Cailliau was a co-author, with Tim Berners-Lee, of the 1990 version of the project proposal to CERN management for the World Wide Web.

He therefore started to write his own "web client" program that would allow the creation, editing, and browsing of hypertext pages. To identify the links in the documents, he developed a simple language called *HyperText Markup Language* (HTML). HTML used labels called *tags* to indicate where there were links to other documents. Publishers had used similar tagging schemes for many years to specify how documents should be formatted for typesetting.

The original HTML suggested by Berners-Lee contained only a dozen tags (see Fig. 11.7). Over the years, new tags were introduced for embedding images, multimedia, and scripts, so that the number of tags is now close to a hundred. To specify document addresses down to the specific file on a specific computer, Berners-Lee came up with the idea of a *universal resource identifier* (URI). The URI consisted of the computer server name followed by the directory path and file name of the document. For example, the address http://info.cern.ch specified the original CERN website, which contained information with links to other documents and sites (Fig. 11.8). The first four letters tell the browser which protocol to use to find the document. Berners-Lee introduced the *hypertext transfer protocol* (HTTP), a set of instructions that specified how a computer could communicate with other computers over the Internet to get to the desired content at a remote

```
<HTML>
        <TITLE>
          A sample HTML instance
        </TITLE>
        <H1>
          An Example of Structure
        </H1>
          Here's a typical paragraph.
        <P>
        <UL>
          <LI>
                Item one has an
                 <A NAME="anchor">
                       anchor
                 </A>
          <LI>
                Here's item two.
        </UL>
</HTML>
```

Fig. 11.7. Example of a simple text with bracketed codes indicating the document elements in HTML from the June 1993 specification.

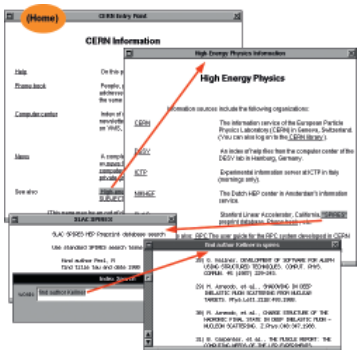


Fig. 11.8. Screenshot from the first text-based browser developed at CERN.

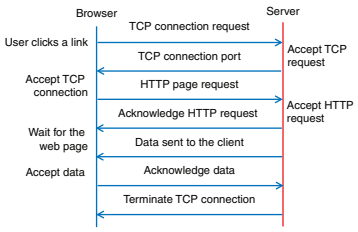


Fig. 11.9. The HTTP is a sequence-response procedure. The figure illustrates the sequence of messages between the browser and the server when the user clicks on a hyperlink.

site. This protocol requests a computer called the *web server* at the remote site to upload the requested web page to the user's browser for viewing (Fig. 11.9).

In November 1990, Nicola Pellow, a student visiting CERN from Leicester Polytechnic (now De Montfort University) in England wrote a web browser called a *line-mode browser* that worked on almost all computer terminals. Her browser enabled Berners-Lee's small team to release the World Wide Web program to people at CERN who had NeXT computers in March 1991. One of the key converts to the web was Paul Kunz, a visitor to CERN from the Stanford Linear Accelerator Center (SLAC) in Palo Alto, California. He was also a NeXT enthusiast. After Kunz returned to the United States, he worked with Louise Addis, the SLAC librarian, and their colleague Terry Hung to make SLAC's catalog of online documents accessible using the web. Berners-Lee announced this offering on 13 December 1991. The first web server in the United States, the SLAC home page, slac.html, was created six months later and was the first web server outside CERN (Fig. 11.10).

To evangelize the World Wide Web, Berners-Lee and Cailliau wrote up their work as a paper for the major Hypertext Conference sponsored by the Association for Computing Machinery that was scheduled to take place in San Antonio, Texas, in December 1991. It is part of the folklore of computer science that their paper was rejected! In spite of this rejection, they asked for permission to give a demonstration. This was not easy to arrange because, at that time, the conference provided no Internet connectivity to attendees. Cailliau had to call up a local university and arrange to use their dial-in service. As Berners-Lee says:

> We were the only people at the entire conference doing any kind of connectivity.… At the same conference two years later … every project on display would have something to do with the Web.[14]

The value of the web increased rapidly as more sites put up web servers and made their content available to other computers using the HTTP protocol. The real breakthrough happened when images started to appear on web pages (Fig. 11.11). In the summer of 1992, David Williams suggested that Berners-Lee should take sabbatical leave from CERN to visit the United States. Berners-Lee spent time at MIT (B.11.6) on the East Coast and at Xerox PARC on the West Coast promoting his ideas. While he was in the San Francisco area, he visited Paul Kunz and Louise Addis at SLAC but also took the time to pay homage to another Bay Area resident, Ted Nelson, who was the original inventor of hypertext.

By the end of 1992, the CERN team had a list of about thirty web servers, mainly in Europe, but also a handful based in the United States. NCSA at the University of Illinois, Urbana, was one of these U.S. sites. Traffic to the first web server at CERN was growing rapidly, with the number of daily *hits* – page views – doubling every three to four months. We can now see the relevance of "Error 404" in the title of this section. The error 404 message appears whenever a web client follows a link to a server but either the server is no longer there or the server is unable to find the requested page. The user typically receives an error message saying "Error 404 – Page Not Found" when attempting to follow such a broken or dead link. Berners-Lee's great insight was to realize that such

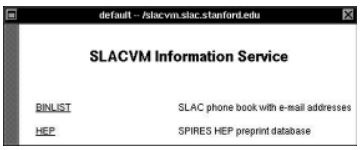
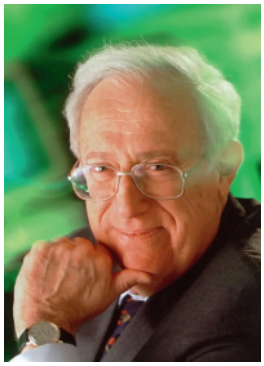


Fig. 11.10. The first U.S. website was set up at SLAC in 1991. Other physics laboratories around the world soon followed their example. These early web servers enabled the physics community to share documents and results of experiments.

Fig. 11.11. A picture of the CERN amateur singing group *Les Horribles Cernettes* was uploaded to the web by their fan Tim Berners-Lee on 18 July 1992, to test an early version of his software. It is believed to be the first photo posted to the web.



B.11.6. Michael Dertouzos (1936–2001) was director of the Laboratory of Computer Science at MIT. He played a crucial role in working with Tim Berners-Lee in establishing the World Wide Web Consortium (W3C) which oversees the development of web standards. In his book *The Unfinished Revolution* he outlined his vision for "human centric computers."

broken links had to be tolerated if the World Wide Web was to scale up and be truly global and dynamic.

The web browser developed by Berners-Lee was specific to the rather uncommon NeXT workstations, and also required new windows to be opened as the user surfed from site to site. Marc Andreessen and Eric Bina at NCSA wanted to give users a much simpler experience and have pictures and text displayed in a single screen as a "mosaic" – that is, an assembly of small pieces of information. Their Mosaic browser was launched in April 1993 and tens of thousands of copies were downloaded from NCSA within the first few weeks (Fig. 11.12). Importantly, NCSA also developed versions of Mosaic for Windows and the PC so that their browser would reach the largest possible audience. From one hundred hits per day in the summer of 1991, by the summer of 1993 the CERN web server was receiving ten thousand hits per day (Fig. 11.13). The exponential growth of the web was now under way in earnest: from about fifty websites at the end of 1992, there were more than ten thousand sites by the end of 1994. The Mosaic browser was an important factor in fueling this growth. Another was CERN's declaration, at the end of April 1993, "that CERN agreed to allow anybody to use the Web protocol and code free of charge, to create a server or browser, to give it away or sell it, without royalty or any other constraint."[15] This declaration was reassuring to industry and paved the way for the development of web-based *e-commerce*, business conducted over the Internet.

After graduating from the University of Illinois, Andreessen moved to Silicon Valley, California's center of high-technology industries, in December 1993. He was keen to find some way to commercialize his work on the Mosaic web browser. With Jim Clark, one of the original founders of the computer manufacturer Silicon Graphics, Inc., Andreessen established the Mosaic Communications Corporation (later Netscape Communications Corporation) (B.11.7). Their first move was to hire the core Mosaic development team from NCSA, and their first browser, Netscape Navigator, was released in December 1994 (Fig. 11.14). Their browser worked with Unix, Windows, and Macintosh systems and was released as a download over the Internet. By making the software free for noncommercial use, Netscape Navigator rapidly became the *default browser* for the web, the browser that launched automatically unless the user changed it.

The dot-com bubble and the browser wars

To understand the rise of e-commerce, it is necessary to understand something about encryption. The science of cryptography dates back to ancient times and consists of techniques for *encoding* the information in a message so that it can only be *decoded* by its intended recipient. We discuss the state of the art in cryptography in more detail in Chapter 12 – here we will assume that reliable and computationally manageable encryption methods exist.

Netscape set the stage for the emergence of e-commerce by providing a new security facility called the *secure sockets layer* (SSL). This offered businesses the option of using an *encrypted* (coded), secure channel for users to send their credit card information over the Internet. The SSL makes the whole business of encryption invisible to the user. By 1994, with Netscape's browser offering SSL

Fig. 11.12. Plaque commemorating the creation of the Mosaic web browser at NCSA at the University of Illinois, Urbana-Champaign.

encryption, all the elements were in place for e-commerce, or online retailing (a.k.a. dot-com), to take off.

To understand the significance of the term *dot-com*, we need to understand the naming convention that determines *domain names* on the Internet. Domain names were introduced in the early days of the ARPANET to serve as easily remembered names for ARPANET resources. The easily remembered domain names corresponded to an unmemorable string of numbers that was the actual *Internet Protocol* (IP) address. IP is the method by which data is sent from one computer to another on the Internet, and an *IP address* is a sequence of numbers that specifies a particular computer on the network.

At first, the mapping of computer host names to numerical addresses was held on a computer in Doug Engelbart's group at the Stanford Research Institute. In 1983, the Internet Engineering Task Force, a group that develops and promotes Internet standards, introduced the Domain Name System, which automatically translates the names we type in our web browser into IP addresses. Today, the Internet Corporation for Assigned Names and Numbers (ICANN) manages the assignment of top-level domain names (Fig. 11.15). A domain name consists of two or more labels, separated by dots, such as *microsoft.com*. The label after the first dot is the top-level domain, in this case, *.com*, signifying a commercial organization.

When the system was devised in the 1980s, there were two main groups of domains – a top-level country domain consisting of a two-letter abbreviation, such as *.uk* for the United Kingdom, and a top-level domain in the United States for seven types of organizations, such as *.com* for businesses. The other six were *.gov* for government, *.edu* for education, *.mil* for military, *.org* for organizations, *.net* for network, and *.int* for international. After the top-level domain comes the second-level domain and so on, as in *southampton.ac.uk*, where the second-level domain name *.ac* (equivalent to *.edu* in the United States) represents an academic organization in the United Kingdom, in this case the University of Southampton. In the domain name www.cern.ch, *www* signifies the web server at the CERN laboratory in Switzerland, which is specified by the *.ch* top-level, country domain name.

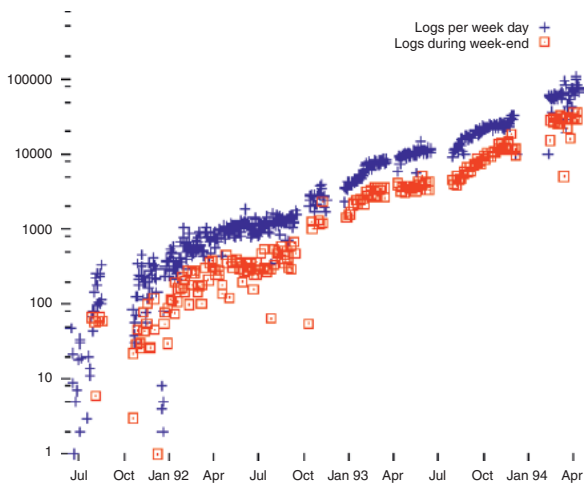


Fig. 11.13. There was rapidly growing interest in the web in the early days from 1992 to 1994. This figure shows the exponentially increasing load on the CERN web server (note the vertical scale).

Fig. 11.14. First release of Netscape Navigator browser in 1994.

Fig. 11.15. The ICANN is a nonprofit organization responsible for assigning top-level Internet domain names.

The first commercial Internet domain name, symbolics.com, was registered in March 1985 by Symbolics, Inc., a computer systems company in Massachusetts. By 1992, fewer than fifteen thousand .com domains had been registered. All this changed in August 1995 when Netscape made an initial public offering (IPO) on the stock market. From an initial stock price of $28 per share, the shares jumped to a peak of $75 on the first day of trading. Netscape was now a company valued at more than $3 billion even though it had yet to generate a profit. Sales for its Netscape Navigator browser were small but growing rapidly each quarter. By the end of 1995, the share price had risen to $175 per share and investors were looking for other .com Internet companies that could show similarly spectacular growth.

One of the earliest and most enduring of the "dot-com" businesses is the online retailer Amazon (Fig. 11.16). Its founder, Jeff Bezos (B.11.8), was an investment analyst in New York when he saw a report in May 1994 describing the explosive growth of the World Wide Web. He recognized the potential of the web for online retailing and decided that selling books online would have a much lower overhead than traditional bricks-and-mortar bookstores. He left his job in Manhattan, moved to Seattle, and established Amazon in July 1994.

When Netscape's IPO started the rush to invest in "dot-coms," Internet start-ups sprung up offering everything from airline tickets and hotel rooms, such as *Lastminute.com*, to selling pet supplies, such as *Pets.com*. Throughout the second half of the 1990s, the dot-com bubble grew. A wave of investor enthusiasm pushed the stock value of Internet companies to greater and greater heights although most of the companies had not yet made any money. These new companies seemingly made a virtue of not having a conventional business plan showing a plausible path to a profitable future! A small group of financial analysts specializing in high-technology companies fueled the dot-com buying frenzy. Having seen the spectacular rise of Netscape, these Internet analysts talked up the value of these dot-com companies to unrealistic values as compared to traditional bricks-and-mortar companies. In February 2000, the inevitable happened: reality returned to the stock market and the shares of dot-com companies went into free fall. Amazon shares, for example, which had reached $600 per share in 1999, fell to less than $10 by the end of 2001.

Many of the dot-com companies disappeared without a trace, but Amazon continued to grow its business and reported a profit by the end of 2002, only one year later than Bezos's original business plan had predicted. What of Netscape, the company that started the boom? Their dominant market share collapsed dramatically when Microsoft Corporation woke up to the potential – and threat – offered by the Internet and the web.

B.11.7. The first CEO of Netscape Jim Barksdale (left) and the two founders, Marc Andreessen and Jim Clark.

The story of Microsoft's conversion to embracing the web and the Internet is long and complicated. One beginning was in September 1991 with the recruitment of James Allard, known within Microsoft as "J Allard." Allard came from a Unix background and believed in the value of open *Application Programming Interfaces* (APIs). An API is a description of exactly how one program needs to ask another software program to perform a specific service. An open API is just an interface whose specification is freely available to the public so that any user can develop applications and services for a particular software program. It was clear to Allard that Microsoft urgently needed

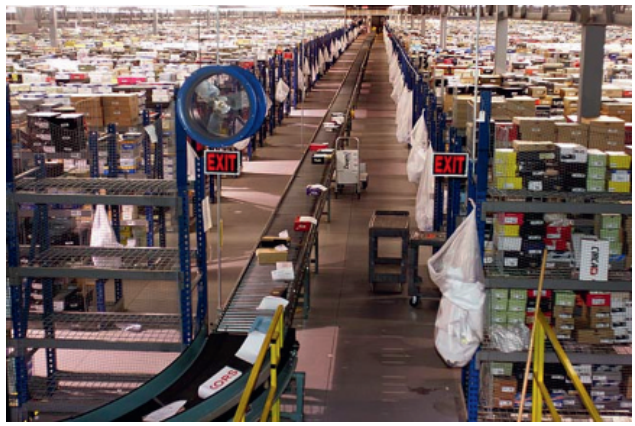


Fig. 11.16. View inside an Amazon warehouse, which sorts products and packs and sends out orders for the online retailer.

an open API to link Windows to the Internet TCP/IP protocols. In the next year or so, Microsoft led a collection of companies in defining this interface, called the *Windows sockets interface* or "Winsock." Microsoft officially endorsed the API in January 1992. Peter Tattam from the University of Tasmania in Australia released an open-source version. His software, called Trumpet Winsock, enabled users to connect to the Internet from Windows 3.0, which had no built-in TCP/IP support. Although few people paid Tattam for his software, Trumpet Winsock helped fuel the growth of the Internet by allowing millions of PC users to connect to the Internet for the first time.

Another step in Microsoft's conversion to the web was caused by a snowstorm in Ithaca, New York. Steven Sinofsky, then Bill Gates's technical assistant, was visiting Cornell University on a recruiting trip in February 1994. A snowstorm shut the airport, and he was forced to return to the campus. Things had changed since he had been a student there only seven years before. Many students now had their own PCs or Macs and campus computing resources could be accessed through the campus TCP/IP network. Students and staff routinely communicated by email, and use of the World Wide Web was growing rapidly. Still trapped in Ithaca, Sinofsky sent an email to Gates headed "Cornell is WIRED."

Twice a year, Bill Gates used to take what he called a "Think Week," spending several days in seclusion reading research papers and pondering the future of technology. Sinofsky's email from Cornell triggered Gates's famous "Think Week" memo in May 1995, which said, "The Internet is a tidal wave. It changes the rules. It is an incredible opportunity as well as incredible challenge."[16] It was a long memo and in it, Gates made many prophetic comments:

> In this memo I want to make clear that our focus on the Internet is crucial to every part of our business. The Internet is the most important single development to come along since the IBM PC was introduced in 1981. It is even more important than the arrival of the graphical user interface (GUI).
>
> The Internet's unique position arises from a number of elements. TCP/IP protocols that define its transport level support distributed computing and scale incredibly well. The Internet Engineering Task Force (IETF) has defined an evolutionary path that will avoid running into future problems even as

B.11.8. Jeff Bezos, founder of Amazon, built the company steadily from its beginnings as a dot-com start-up in 1994 to its current position as the dominant, global, online retailer. He was inspired to resign from his job as an investment analyst in New York after seeing the explosive growth of the World Wide Web.

Fig. 11.17. The "Browser Wars" were a rivalry between Netscape Navigator and Microsoft's Internet Explorer. As a joke, in the early hours of the morning, the Microsoft team celebrated the launch of IE 4.0 in 1997 by placing a giant e on the lawn in front of Netscape's buildings in Mountain View, California.

> eventually everyone on the planet connects up. The HTTP protocols that define HTML Web browsing are extremely simple and have allowed servers to handle incredible traffic reasonably well. All of the predictions about hypertext – made decades ago by pioneers like Ted Nelson – are coming true on the Web.
>
> Amazingly it is easier to find information on the Web than it is to find information on the Microsoft Corporate Network. This inversion where a public network solves a problem better than a private network is quite stunning.
>
> I think that virtually every PC will be used to connect to the Internet and that the Internet will help keep PC purchasing very healthy for many years to come.[17]

As a result of Gates's memo, Microsoft dramatically changed course. In August 1995, at the same time as the company launched the Microsoft Network (MSN), a collection of Internet sites and online services, it also included a web browser called *Internet Explorer* in its release of *Windows 95*. It soon became clear that the vast majority of users preferred connecting to the free Internet rather than subscribing to a commercial consumer network like MSN or America Online (AOL).

Microsoft's first browser, based on licensing the original NCSA Mosaic code, was fairly primitive. Over the next two years, Internet Explorer and Netscape Navigator battled for supremacy in the browser market, with each company releasing several browser upgrades each year (Fig. 11.17). By January 1998, Internet Explorer was not only the technical equal of Netscape Navigator but was also considerably more stable, with many fewer bugs. Because Internet Explorer was bundled free with the Windows operating system, it rapidly became the browser of choice for PC users. Netscape's share price fell from its peak of $175 per share to under $15.

Before AOL acquired Netscape in 1999, Netscape released the source code for the browser and set up the Mozilla Foundation to manage its future development. The foundation describes itself as "a non-profit organization that promotes openness, innovation and participation on the Internet."[18] In 2004 Mozilla released the Firefox browser and, by 2007, it had gained a significant market share, despite the dominance of Internet Explorer and new entrants to the browser world, such as Safari from Apple and Chrome from Google.

Fig. 11.18. An artist's impression of the millions of web pages.

Internet search and the PageRank algorithm

By the mid-1990s, the number and types of websites on the World Wide Web had grown enormously. In the early days of the web, users had to find out about "good" websites by word of mouth. Now, finding specific information on the web was like looking for a needle in a haystack – a user needed to search through an unorganized jumble of millions of web pages (Fig. 11.18). In 1994 two Stanford graduate students, Jerry Yang and David Filo (B.11.9), created a website called "Jerry's Guide to the World Wide Web," which was an alphabetical directory of interesting websites. Later that year, they renamed the website Yahoo! and set up a company that grew rapidly during the dot-com boom. In an attempt to keep up with the rapid growth of the web, Yahoo! employed teams of editors to help select the websites for the directory.

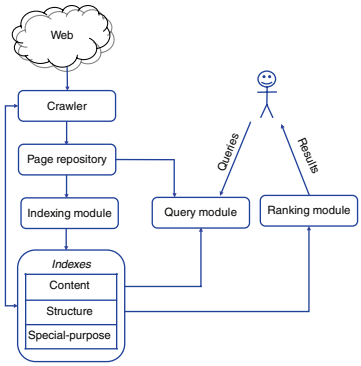


Fig. 11.19. Basic structure of a search engine showing the query-independent elements that respond to a user's query.

Other companies took a different approach by providing web *search engines*. These companies attempted to supply an index to the content of web pages at the different sites. A search engine works just like the index in a book in helping the reader look up a particular topic. By 1998 the leading search engine, with more than 50 percent market share, was AltaVista. Computer scientist Paul Flaherty at Digital Equipment Corporation's (DEC's) Network Systems Laboratory in Palo Alto had the idea of DEC building a web index. He recruited colleagues Louis Monier and Michael Burrows to write the software for what became the AltaVista search engine.

To create indexes for individual web pages, a search engine must first search out and capture these web pages (Fig. 11.19). This search is done with a *web crawler*, a piece of software that follows hypertext links to discover new web pages. The crawler sends out "spiders," which are given explicit instructions on where to start crawling and what strategy to use in following links to visit new pages.

The web pages returned by the spiders now need to be indexed. The indexing software takes each new page, extracts key information from it, and stores a compressed description of the page in one or more indexes. The first type of index is called the *content index*. This directory stores information about the different words on the page in a structure known as an "inverted file," which is similar to the index in the back of a book. Next to each term being indexed, the inverted file keeps information, such as the page numbers on which the term appears. We can now do single-word queries to find the relevant web pages. Of course, to efficiently handle more complex queries, we need to store more than just the page number for each word. We can add extra information, such as the number of times a word appears on a page, its location on the web page, and so on. A key advance made by AltaVista was also to store information about the HTML structure of the web page. By looking at the HTML tags on the page, we can identify whether the word being queried appears in the title, in the body of the page, or in the *anchor text*, the specific word or words used to represent the hypertext link. All of this indexed information is combined to deliver an overall "content score" for each web page to determine the most relevant page in answer to a query. It was this combination of content and structure information that had made AltaVista the leading search engine by 1998.

Modern search engines use more than just the content and structure to determine the best websites to return in answer to a user's query. It was the development of the *PageRank algorithm* by two Stanford graduate students,

B.11.9. David Filo and Jerry Yang founded the search company Yahoo! Inc. as Stanford graduate students. In 1994 they started compiling a directory of websites and extended the portal with a range of online services. By 1996 the company went public and became one of the landmark successes of the dot-com era. After the dot-com crash in 2000, the company suffered significant losses but Yahoo remains one of the household names of the Internet age, delivering online services to millions of customers.

Fig. 11.20. Aerial view of Google headquarters, the Googleplex, in Mountain View, California. The roofs on the buildings are covered with solar panels.

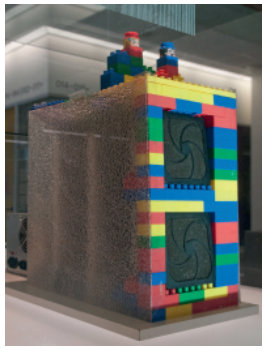


Fig. 11.21. Larry Page and Sergey Brin's first server at Stanford was encased in Lego blocks.

B.11.10. Eric Schmidt, Sergey Brin, and Larry Page (left to right) shown answering questions in 2008. While PhD students at Stanford, Sergey and Larry came up with the idea of ranking the web pages based on their link structure. They described their ideas in a much-cited paper "The Anatomy of a Large-Scale Hypertextual Web Search Engine." After unsuccessfully trying to sell their ideas to AltaVista and Yahoo!, the two leading web search companies at that time, they set up Google, beginning in the traditional Silicon Valley garage.

Sergey Brin and Larry Page (B.11.10), that formed the basis for the success of Google (Fig. 11.20). The PageRank algorithm was a step-by-step procedure to calculate an "importance score" for each web page. Instead of just looking at the content and structure of web pages, Page and Brin analyzed the hypertext link structure. By combining their importance score from the link analysis with the content score from traditional indexes, Brin and Page developed a search engine that almost magically delivered the most useful websites to users.

Larry Page was intrigued by AltaVista's information about the hypertext links and decided that analyzing the structure of the link data could be valuable. To do this, he started downloading as much as possible of the entire World Wide Web to his computer (Fig. 11.21). Meanwhile, Sergey Brin with his Stanford adviser, Rajeev Motwani, had been investigating the currently available search engines and directories. Page and Brin then teamed up on trying to accomplish Page's goal of downloading as many web pages as possible and analyzing their link structure. Coming from an academic family, Page had the idea that the number of links to a web page was similar to the citation count of a scientific paper. The number of times other authors cite a paper is a significant indicator of the importance of the research described in the paper. However, Page realized that just counting the number of links pointing to a web page does not give the full measure of the importance of the page. Just as citations to a scientific paper from Nobel Prize recipients are more significant than citations by ordinary mortals, so too were links to a web page coming from an important or authoritative site. The journalist and author David Vise describes Page's idea as follows:

> All links were not created equal. Some mattered more than others. He would give greater weight to incoming links from important sites. How would he decide what sites were important? The sites with the most links pointing to them, quite simply, were more important than sites with fewer links.[19]

In a play on his last name, Page called his new algorithm PageRank.

How do you calculate the PageRank score of a given web page? If we assign an initial "authority" of 1 to each web page, we can calculate the accumulated authority of a given page by adding up the authorities of all the web pages that point to it. Unfortunately, the graph of web page links may contain a "cycle" – that is, by clicking on web links you eventually get back to the starting point (see Fig. 11.22). This makes it impossible to calculate an authority score for the sites in the cycle. To avoid this problem, Page and Brin introduced a "random surfer." Imagine a surfer roaming the web following

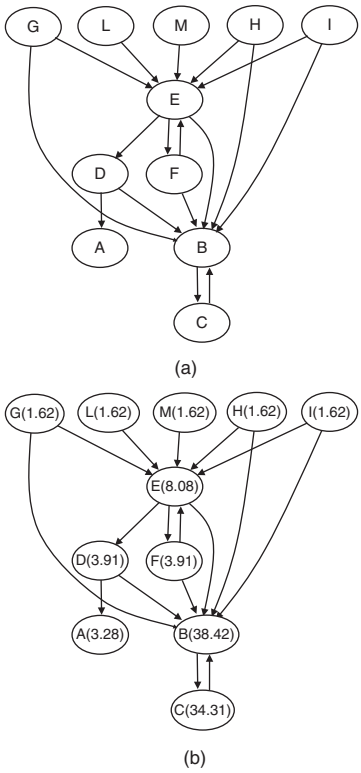


Fig. 11.22. (a) An example of a set of web pages – *nodes* – showing the hypertext links connecting the different pages with the direction of each link marked by an arrow. Web page A is known as a *dangling node* because this has no outgoing links. Pages B and C are said to form a *bucket*, a reachable strongly connected pair of pages with no outgoing links to the rest of the pages of the graph. (b) The importance of each page can be calculated using the PageRank algorithm. For this graph we have calculated the importance of each page using the *random surfer* method although in practice other methods are used. The scores have been normalized to add up to 100. Page C receives just one link but this is from the most important page. Page C ends up being more important than page E – which receives more links but all but one from pages G, H, M, L, and I, which all have no incoming links and therefore have the minimum importance.

hypertext links from page to page. When he or she arrives at a web page with multiple links, the surfer randomly chooses one of the links. We can now attempt to calculate the importance of each page by counting how many times the random surfer visits each site. To avoid the problem of *buckets* – cycles where the surfer can go around forever – Page and Brin introduced a "teleportation" probability. At each page the surfer visits, there is a chance that he or she does not follow one of the links from the page but instead jumps to an entirely new web page chosen at random. The surfer then uses this page as a new starting point for continued surfing. In addition, to avoid the problem of the surfer getting stuck in a so-called *dangling node*, a page with no outgoing web links, Page and Brin also allowed the surfer to jump to a random new page from such nodes. We can simulate the action of the random surfer on a network of web links and find out which pages the surfer visits most often. The importance score calculated in this way takes into account both the number of hyperlinks pointing to a web page and the authority of the linking web page. The method works in the presence of link cycles and dangling nodes and is able to deliver meaningful and stable values for each page's importance.

Figure 11.22a shows a network of eleven web pages and their associated links. A computer simulation of the random surfer algorithm gives the PageRank score shown in Figure 11.22b. In practice, the computation of the PageRank score is not done by performing a computer simulation of the surfer's travels. Instead, the problem can be formulated as solving a mathematical problem involving *sparse matrices*. A *matrix* is an array of numbers or symbols arranged in rows and columns, and *sparse* just means that the matrix has many elements that are zero. The matrices concerned are huge, with many billions of rows, but fortunately fast computational methods exist that can be used to find the PageRank scores.

Armed with the PageRank algorithm, by early 1997 Page had developed a prototype search engine that he called "BackRub" because it analyzed the incoming or "back" links to web pages to calculate a page's importance. While other searches relied on the content and structure indexes, BackRub added the dimension of importance to rank pages in order of likely relevance. By the end of the year, Page and his officemate at Stanford, Sean Anderson, were brainstorming for a new name for the search engine. They decided on the name Google, thinking the word *google* meant a very large number. Page registered Google.com the same evening only to find out next day that the word they were thinking of was spelled *googol*, meaning the number 1 followed by one hundred zeros. With their search engine now being used by Stanford students and faculty, Page and Brin needed to beg and borrow as many computers as they could to keep up with both the growth of users and that of the web. Page's dorm room became the first Google data center, crammed with inexpensive PCs (Fig. 11.23). Today Google has data centers around the world that run hundreds of thousands of servers (Fig. 11.24).

Having demonstrated the superiority of a search engine using the PageRank algorithm to calculate a web page's importance, Page and Brin tried to interest AltaVista in buying the rights to their system. Paul Flaherty, one of the originators of the AltaVista search engine, was impressed with their link-based approach to page ranking. Nevertheless, Flaherty came back to them saying

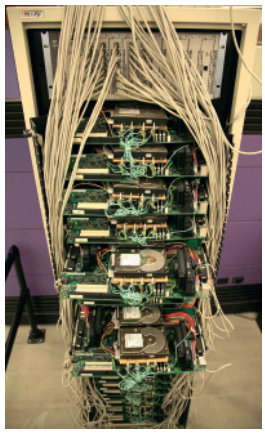
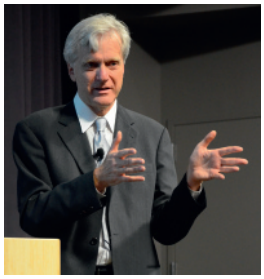


Fig. 11.23. The first Google production server rack from about 1998. The rack has multiple computer boards stacked one above the other.



B.11.11. Andy Bechtolsheim, one of the founders of Sun Microsystems, was Google's first investor.

that Digital was not interested. Page and Brin got the same lack of interest from other search-engine companies and from Yahoo!. However, Yahoo! cofounder David Filo gave them some valuable advice: take a leave of absence from Stanford and start their own business. By the summer of 1998, Page and Brin had decided to take Filo's advice. They had a meeting with Andy Bechtolsheim (B.11.11), one of the original founders of Sun Microsystems and now a Silicon Valley investor in start-up businesses. Despite his concerns about the lack of a viable business model for search engines in general, Bechtolsheim was impressed by Page and Brin and, without any discussion of stock distribution, he wrote them a check for $100,000, made out to Google Inc. Page and Brin had to keep the check uncashed for two weeks until they had incorporated Google as a company and opened a bank account.

Google's rapid rise to dominance in web searching is chronicled in detail in *The Google Story* (Fig. 11.25) by David Vise. A company called Overture had pioneered search-related advertising, providing one of the first business models for search engines. Despite their initial reluctance to embrace an advertising model, Page and Brin decided to implement a variant of Overture's basic idea. To maintain the integrity of their free search service, they insisted on keeping their home page entirely free from advertisements and on distinguishing the free search results from what they called "sponsored links." Advertisers bid in an online auction for priority placement in these sponsored links, which appeared when users searched on specific search terms. Google only made money when a user clicked on one of the ads displayed. By the year 2000, Google was handling fifteen million searches per day compared to ten thousand only eighteen months before.

Google's early breakthrough undoubtedly owed much to the idea of including an importance score based on PageRank. However, *search engine optimization* (SEO) companies have now become big business. These SEO companies advise advertisers how to ensure that their web pages will appear near the top of search results. Some SEOs offer more than just good advice about how to increase the advertiser's content and importance score. They also try to manipulate the results using *web spam*, the practice of manipulating web pages

Fig. 11.24. One of the Google Data Centers with tens of thousands of computers installed on racks. The blue light indicates that the servers are running well.

Fig. 11.25. The cyber power stations of the twenty-first century. Steam is rising from the cooling towers of Google Data center in "The Dalles, Oregon."

deceptively so that search engines rank the pages higher than they would without any manipulation. Web spam can take a number of forms. *Term spamming* employs the term in question repeatedly on the web page, sometimes using white text on a white background so that the spam is invisible to human readers. *Cloaking* involves deceiving a web crawler by having different pages for normal requests than for spiders. The spider returns a clean web page without spam. Another way to influence the importance score is to automate the creation of a large number of different web pages with links to the customer's site. For all these reasons, modern search engines now include many other factors besides PageRank in arriving at their ranking of search results. For example, Harry Shum from Microsoft's Bing search engine team has said, "We use over 1,000 different signals and features in our ranking algorithm."[20]

The social web and beyond

B.11.12. Ward Cunningham invented the first collaborative *wiki*, a website that allows users to make changes and add their own contributions. Wiki is a Hawaiian word meaning fast or quick.

Darcy DiNucci, a consultant on information design, first introduced the term *Web 2.0* in 1999 in an article called "Fragmented Future":

> The relationship of Web 1.0 to the Web of tomorrow is roughly the equivalent of *Pong* to *The Matrix.* Today's Web is essentially a prototype – a proof of concept. This concept of interactive content universally accessible through a standard interface has proved so successful that a new industry is set on transforming, capitalizing on all its powerful possibilities. The Web we know now, which loads into a browser window in essentially static screenfuls, is only an embryo of the Web to come.
>
> The first glimmerings of Web 2.0 are beginning to appear, and we are just starting to see how that embryo might develop.… The Web will be understood not as screenfuls of text and graphics but as a transport mechanism, the ether through which interactivity happens.[21]

The term *Web 2.0* was later popularized by computer book publisher Tim O'Reilly at the first Web 2.0 conference, held in 2004. The term does not refer to an update of any particular technical specification but rather to the way in which software developers and users are now using the web, focusing on collaboration, user-generated content, and social networking. Web 2.0 applications allow users to interact and collaborate in new ways to create virtual communities. This development is visible through the growth in popularity of the online journals called *blogs* and of *wikis*, websites that allow users to make changes and add their own contributions.

The word *blog* is an abbreviation of *web log* and is usually an online diary recording the thoughts or actions of an individual (Fig. 11.26). Blogs are interactive in that, on most blogs, readers can leave comments and participate in an online discussion. The growth of blogging was accelerated by new, easy-to-use web publishing tools that did not require the user to have any knowledge of technologies such as HTML or FTP. By 2011, there were more than 150 million public blogs.

A *wiki* is a website that allows its users to interact with the site to add, modify, or delete content. The name *wiki* is a Hawaiian word meaning *fast* or *quick*. Ward Cunningham (B.11.12), inventor of the wiki, described it as "the simplest

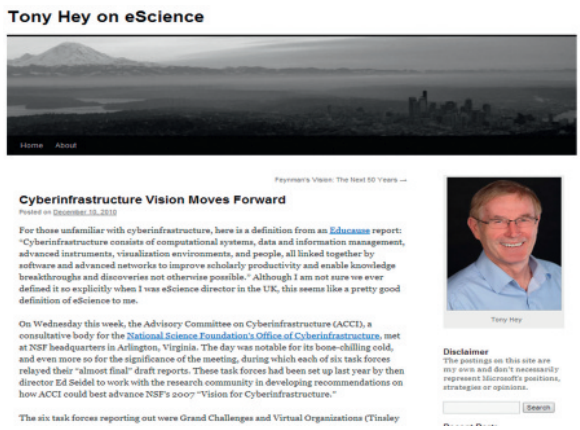


Fig. 11.26. Tony Hey's personal blog on e-Science.

online database that could possibly work."[22] He produced software that allowed users to interact with the wiki using a web browser. In some ways it is remarkable that public wikis work at all. For example:

> Most people, when they first learn about the wiki concept, assume that a Web site that can be edited by anybody would soon be rendered useless by destructive input. It sounds like offering free spray cans next to a grey concrete wall. The only likely outcome would be ugly graffiti and simple tagging, and many artistic efforts would not be long lived. Still, it seems to work very well.[23]

Wikis do sometimes get vandalized, but most users abide by the rules chosen collaboratively for the wiki's governance. The most famous wiki of all is, of course, Wikipedia (B.11.13), the online encyclopedia created by its users. It sums up its policies and guidelines in five pillars:

- Wikipedia is an encyclopedia
- Wikipedia has a neutral point of view
- Wikipedia is free content
- Wikipedians should interact in a respectful and civil manner
- Wikipedia does not have firm rules

Although it undoubtedly has its faults, Wikipedia is a remarkable collaborative creation and makes a massive amount of content freely available.

Another important attribute of Web 2.0 is the ability of users to invent their own tags and to use these to bookmark photos and other material on the web. For example, Flickr allows users to tag their photos and use these tags for organizing and searching through their collection. Tagging is also used by virtual communities to create *folksonomies*, a grassroots way of classifying content based on user-generated tags or keywords that annotate and describe the information. Unlike a traditional hierarchical taxonomy, in a folksonomy all tags have more or less equal status. Finally, Web 2.0 offers the capability to create *mash-ups*, web pages or applications that allow users to combine data from multiple websites. It is common to make mash-ups with map data and to overlay different data sets such as houses for sale, the traffic status, and so on.

B.11.13. Jimmy Wales worked in the finance industry before starting the free web encyclopedia Wikipedia in 2001. *Time* magazine named him in its list of "The 100 Most Influential People in the World" in 2006.

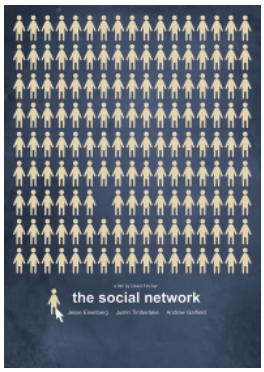


Fig. 11.27. The founding of Facebook was made into a movie called *The Social Network*. The marketing slogan for the movie was "You don't get to 500 million friends without making a few enemies."

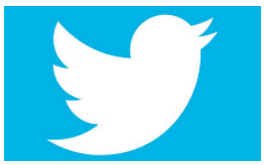


Fig. 11.28. The Twitter service was introduced in 2006 and in 2013 had more than half a billion users. Twitter is an Internet text-messaging service that allows a maximum message size of 140 characters.

From these beginnings, new companies such as Facebook (B.11.14) and Twitter have now emerged. Facebook is a social networking site whose users can share personal updates, photos, and other information with their friends (Fig. 11.27). Twitter is a microblogging service that lets a person send brief text messages up to 140 characters in length to a list of followers. Twitter's celebrity "tweeters" can attract millions of followers (Fig. 11.28).

Berners-Lee disagrees that Web 2.0 constitutes an essentially new vision for the web and dismisses it as marketing jargon:

> Web 1.0 was all about connecting people. It was an interactive space, and I think Web 2.0 is, of course, a piece of jargon, nobody even knows what it means. If Web 2.0 for you is blogs and wikis, then that is people to people. But that was what the Web was supposed to be all along.[24]

Instead, Berners-Lee is looking toward what he calls a "Semantic Web" in which machines can process and understand the actual data on the web:

> The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation. The first steps in weaving the Semantic Web into the structure of the existing Web are already under way. In the near future, these developments will usher in significant new functionality as machines become much better able to process and "understand" the data that they merely display at present.[25]

Industry has taken the first small steps toward such a semantic future. The search-engine companies Google, Microsoft, Yandex, and Yahoo! have agreed on a common vocabulary with which website managers can mark up their sites so that search engines can understand the nature of the content (schema.org). By inserting *microdata* – a set of machine-readable tags introduced with HTML5, the fifth revision of HTML – into the web page, and by using the agreed vocabulary, the website can specify the type of content on the site. For example does the website refer to *Casablanca* the movie or Casablanca the place? The search engine can now distinguish between the sites and better address the user's query. These improvements are all part of the need to go beyond searching and better understand the user's intent.

B.11.14. Mark Zuckerberg was still an undergraduate student at Harvard when he came up with the idea of using the web to link together a network of friends with a social networking website. This modest beginning has now evolved into the social networking company Facebook.

Key concepts

- Hypertext links
- World Wide Web: HTTP, HTML, and URIs
- Web browsers
- Domain names
- Internet search and PageRank
- Web 2.0 and the Semantic Web

Markup language goes hyper

During the twentieth century, typesetting advanced from "hot metal" to "cold type" composition, with photographic negatives replacing metal type as the source for making printing plates. By the 1960s, computer-driven phototypesetting machines had become commonplace, and it was against this background that in 1967, William Tunnicliffe, chairman of the Graphic Communications Association, proposed that a standard set of editorial markup instructions should be developed that could be inserted into a manuscript as directions to typesetters for printing. Charles Goldfarb at IBM adapted Tunnicliffe's idea to develop a business system that would solve the problems of law firms in creating, editing, and printing documents. With his colleagues Edward Mosher and Raymond Lorie, Goldfarb created the *Generalized Markup Language* (GML) in 1973. GML was a set of rules and symbols that described a document in terms of its organizational structure and its content elements and their relationship. GML markups or tags described such parts of a document as chapters, important sections, paragraphs, lists, tables, and so on. By October 1986, the International Organization for Standardization (ISO) had adopted this language as an international standard called the *Standard Generalized Markup Language* (SGML). (The ISO is an international body that attempts to establish uniform sizes and other specifications to ease the exchange of goods between countries.) Berners-Lee wanted to make his new hypertext scheme as simple as possible but, at the same time, keep the goodwill of the global documentation community. He therefore deliberately designed HTML to look like a subset of SGML with only a small set of tags, inserted between angle brackets, as in <word>. Although Berners-Lee never thought that people would use a browser/editor to actually write web pages, the readability of HTML meant that many people did start writing their own HTML documents directly.

Emoticons

An emoticon is a pictorial representation of a facial expression that is meant to indicate the user's mood at the time. The word is a portmanteau word made from the English words *emotion* and *icon*. In the online world emoticons are made up from regular keyboard characters such as :-) and :-(for happy and sad emotions of the sender. They are now often replaced by small images corresponding to these emotions such as ☺ and ☹. The use of emoticons can be traced back to the nineteenth century. The 1857 edition of the National Telegraphic Review and Operators Guide recorded the use of the number 73 in Morse code to express "love and kisses."

Digital forms of emoticons were first proposed by Scott Fahlman to distinguish serious posts from jokes. His email to his colleagues read:

> 19-Sep-82 11:44 Scott E. Fahlman :-)
> From: Scott E. Fahlman<Fahlman at Cmu-20c>
> I propose that the following character sequence for joke markers:
> :-)
> Read it sideways. Actually, it is probably more economical to mark things that are NOT jokes, given current trends. For this, use
> :-(

The actual "smiley face," with two black dots for eyes and a black upturned curve for a mouth, both on a yellow circular background, was created by the artist Harvey Ball in 1963 (Fig. 11.29).

Fig. 11.29. (a) The computer scientist Scott Fahlman and his smiley emoticon. (b) The first smiley from 1963.

The sock puppet and the dot-com bubble

The fate of Pets.com is now a textbook lesson in the need to have a business plan that is grounded in reality. Pets.com was a company whose business was selling pet accessories and supplies to consumers directly over the web (Fig. 11.30). The company was launched in August 1998 and went public on the NASDAQ stock exchange in 1999. Despite the fact that its revenues were less than $1 million in 1999, it spent nearly $12 million of its start-up funding on a high-profile advertising campaign. This included a popular Pets.com sock puppet that was interviewed on the television show *Good Morning America* and an expensive TV commercial that ran during the 1999 Super Bowl. In the dot-com crash, Pets.com stock fell from more than $11 per share in February 2000 to $0.19 by 6 November 2000, the day it closed its doors and sold its assets to pay its debts.

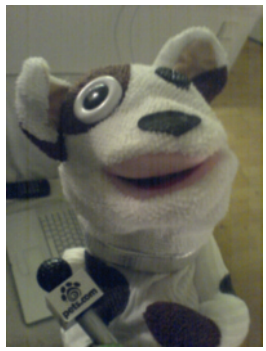


Fig. 11.30. The Pets.com sock puppet was symbolic of the dot-com bubble. The company spent more than ten times its annual revenue on advertising and their popular sock puppet was even interviewed on the prime time TV show *Good Morning America*. The company went from a successful IPO to liquidation in less than a year.