ECONSTOR
Make Your Publications Visible.

A Service of
ZBW
Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Vossen, Gottfried; Hagemann, Stephan

Working Paper

From Version 1.0 to Version 2.0: A brief history of the web

ERCIS Working Paper, No. 4

Provided in Cooperation with:
European Research Center for Information Systems (ERCIS), University of Münster

Suggested Citation: Vossen, Gottfried; Hagemann, Stephan (2007) : From Version 1.0 to Version 2.0: A brief history of the web, ERCIS Working Paper, No. 4, Westfälische Wilhelms-Universität Münster, European Research Center for Information Systems (ERCIS), Münster

This Version is available at:
https://hdl.handle.net/10419/58411

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

Working Papers

ERCIS – European Research Center for Information Systems
Editors: J. Becker, K. Backhaus, H. L. Grob, T. Hoeren, S. Klein,
H. Kuchen, U. Müller-Funk, U. W. Thonemann, G. Vossen

Working Paper No. 4

From Version 1.0 to Version 2.0: A Brief History of the Web

Gottfried Vossen, Stephan Hagemann

ISSN 1614-7448

cite as: Gottfried Vossen, Stephan Hagemann: From Version 1.0 to Version 2.0: A Brief History Of the Web. In: Working Papers, European Research Center for Information Systems No. 4. Eds.: Becker, J. et al. Münster 2007.

Content

1 Introduction 3

2 A New Breed of Applications: The Rise of the Web 5
2.1 The Arrival of the Browser 5
2.2 The Flattening of the World 9
2.3 From Linking to Searching 11
2.4 Commercialization of the Web 19
2.5 P2P Networks and the No-Cost Mentality 28

3 Technology Advances in a Nutshell 32

4 User Perception and Participation: Acceptances of the Web 41
4.1 Blogs and Wikis 41
4.2 Social Networks 48

5 Merging the Streams: The Arrival of “Web 2.0” 53

References 56

List of Figures

Figure 1: NCSA Mosaic. 6

Figure 2: The Client/Server Principle. 7

Figure 3: Netscape Navigator 4. 8

Figure 4: Navigation through the Web along hyperlinks. 12

Figure 5: Anatomy of a Search Engine. 15

Figure 6: Yahoo! Mindset search result for term “bank.” 16

Figure 7: The Long Tail (of Search Queries). 17

Figure 8: The dmoz.org example of a portal page. 19

Figure 9: The Yahoo! home page as of November 2006. 19

Figure 10: Intershop Enfinity Suite 6. 23

Figure 11: Secure payments through a third party. 27

Figure 12: File Request in a Gnutella Network. 30

Figure 13: Home Broadband Penetration in the US. 35

Figure 14: Cient-side vs. Server-side scripting. 37

Figure 15: AJAX and its constituents. 38

Figure 16: The Web service paradigm. 38

Figure 17: Bob Lutz’s blog on auto news from Detroit. 43

Figure 18: Software Design Pattern Wiki in presentation mode. 45

Figure 19: Software Design Patterns Wiki in editing mode. 45

Figure 20: Bibliographic details for Wikipedia entry „Wiki“. 47

Figure 21: MySpace.com Home Page. 51

1 Introduction

The Web has revolutionized our lives and world since its inception a little over a decade ago, more than many other developments in recent history. However, during the year 2006, everything, including the Web, seems to have come out in Version 2, Release 0 (commonly abbreviated as "2.0"): Charles Winkler directed *The Net 2.0*, a sequel to the movie featuring Sandra Bullock that was directed by his father Irwin Winkler in 1995 and that dramatizes the total loss of a person's identity, given the fact that all relevant private information can be identified, in one way or another, on the Internet. The IEEE Spectrum journal reported on *Silicon Valley 2.0 (beta)* in its August 2006 issue, which essentially refers to the rebirth of the Internet hype we have already seen at the end of the last century. In September, the German tech news ticker heise.de reported that "technology keeps the *Family 2.0* together" (details, for our German readers, are at www.heise.de/newsticker/meldung/78809), the basic message being that telecommunication and media are more and more employed these days to accomplish family tasks and to even keep families together. The Australian edition of the T3 tech magazine reported on *Gadgets 2.0* in its October 2006 issue, referring to the next-generation of play stations, intelligent cell phones, music and video players that are currently entering the marketplace. The San Jose Mercury News reported on *India 2.0* in early December. Heise even took it as far as reporting about the "Pub 2.0" (www.heise.de/newsticker/meldung/74527), a mash-up combining location information with comment and search functions, so that users can people can get in contact with each other, for example, via their favourite pub. There are many other such examples, including *Jobs 2.0, Health 2.0, Entertainment 2.0,* or *Music 2.0*.

All of this can easily be traced back to Tim O'Reilly and O'Reilly Media, where the term was coined in late 2004. As can be read at http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html, "the concept of 'Web 2.0' began with a conference brainstorming session between O'Reilly and MediaLive International. Dale Dougherty, web pioneer and O'Reilly VP, noted that far from having 'crashed,' the web was more important than ever, with exciting new applications and sites popping up with surprising regularity. What's more, the companies that had survived the collapse seemed to have some things in common. Could it be that the dot-com collapse marked some kind of turning point for the web, such that a call to action such as 'Web 2.0' might make sense? We agreed that it did, and so the Web 2.0 Conference was born."

"Web 2.0" has rapidly become a label that everybody seems to be able to relate to: The Information Resource Center at the International University in Bremen, Germany, was looking for a "Web 2.0 / Library 2.0 Developer" in an ad in the c't magazine (Issue 23, 2006). Hardware manufacturer Intel suddenly entered the software market and created an Internet Business Suite called "SuiteTwo," an integrated package of Web 2.0 software. US marketing research giant Gartner recognized a major driver for the IT industry in Web 2.0 technologies and awarded it the "Hype Oscar" of 2006. SEOmoz, a Seattle-based search engine optimization company, even rated and ranked over 300 "Web 2.0 companies" in 38 categories and gave a "Web 2.0 Award" 2006 in 21 of them (http://www.seomoz.org/web2.0/). In light of this, it is surprising that an article by Michael M. Roberts entitled *Lessons for the Future Internet – Learning from the Past* (Educause, July/August 2006, pp. 16-24) can do entirely without using that fancy term. And critics are already warning of a "Bubble 2.0," while the New York Times jumps to a "Web 3.0" in November 2006.

But what actually is “2.0” in the context of the Web, what is new and what is not? The media have tried to answer this question in recent past through an exploding number of blogs, show cases, magazine and journal special issues, the famous O’Reilly conferences, and news messages which have offered numerous “definitions” and explanations of Web 2.0, what its features are, what can be expected from it in the very near future, and how important it is to join the bandwagon. Among other effects, this has made many people and companies nervous. They now ask themselves questions like “are we Web 2.0 ready?” or “do we comply to “Web 2.0?” or “what will happen to us if we do not convert to this movement?” Even though the O’Reilly Radar recently published “Web 2.0 Principles and Best Practices,” the situation remains unclear, since even the “eight core patterns” of Web 2.0 identified in that report do not, at closer look, provide a precise definition. In other words, we are looking at a moving target characterized by a certain amount of fuzziness that lacks a simple and concise description and impact analysis.

The goal of this paper is to put the Web 2.0 phenomenon in perspective by taking a tour of the history of the Web, during which we try to identify three major streams of development and impact:

- The *application stream* is simply an account of what we have experienced over the past 10 to 15 years, namely the Web as an ever-growing and omnipresent library of information which we access through search engines and portals, the Web as a commerce platform through which people and companies nowadays do major portions of their business, and the Web as a media repository that keeps lots of things around for free.

- The *technology stream* will touch upon the technological advances underneath the Web, which have enabled its evolution and development (and which actually keep evolving).

- The *user perception and participation stream* will look at how people perceive and use the Web, and how this has changed over the past 15 years in some considerable ways. Since 15 years is roughly half a generation, it will not come as a surprise that especially the younger generation nowadays deals and interacts with the Web in an entirely different way than people did when it all started.

Taken these streams, their impacts, and their results together, we will how the Web consistently evolved from its Version 1.0 and will arrive what is currently considered its Version 2.0. The point will be that Web 2.0 is *not* a new invention of some clever business people, but it is the most recent consequence and result of a development that has started over 10 years ago. Indeed, we will identify three major dimensions along which the Web in its 2.0 version is evolving – data, functionality, and socialization. The reader is referred to [18] for one of the sources that has triggered the discussion we are trying to contribute to, and the fuzz we will try to clean up a little.

The paper represents a preliminary version of Chapter 1 of [28], so feedback and comments from our readers are more than welcome.

2 A New Breed of Applications: The Rise of the Web

Imagine back in 1993, when the World-Wide Web, the WWW, or the *Web* as we have generally come to call it, had just arrived. Especially in academia, where people had been using the Internet since the late 1970s and early 1980s in various ways and for various purposes including file transfer and email, it quickly became known that there was a new service around on the Internet. Using this new service, one could request a file written in a language called HTML, and if one had a program called a "browser" installed on his or her local machine, that program was able to display or "render" the HTML file when it arrived. Let us start our tour through the history of the Web by taking a brief look at browsers and what they are about.

2.1 The Arrival of the Browser

NCSA Mosaic

An early browser was *Mosaic*, developed by the National Center for Supercomputing Applications (NCSA) at the University of Illinois in Urbana-Champaign. There had been earlier browser developments (e.g., Silversmith), but Mosaic was the first *graphical* browser which could display more than just plain ASCII text (which is what a text-based browser does). As can be read on the NCSA Web site (http://www.ncsa.uiuc.edu/News/MosaicHistory), the first version of Mosaic had the following capabilities:

Native support for accessing documents and data using World Wide Web, gopher, Anonymous FTP, and NNTP (Usenet News) protocols. Support for archie, finger, whois, and Veronica (as well as others) through gateways.

- Full HTML display
- Support for in-lined GIF images in HTML hypertext documents
- Internal sound support for Sun/NeXT/DEC .AU audio files
- Support for GIF, JPEG, MPEG, QuickTime for Windows, Microsoft Video for Windows, Postscript and other documents via forking to appropriate viewers (user configurable)
- Full hypertext support, including using in-lined images as anchors
- Optional toolbar with shortcut buttons and optional status bar to display hyperlink destinations
- Local History
- Fully customizable font selection
- Saveable preferences, including window size and position
- Minimal Hotlist

- Scrolling for large documents
- Status of loading/decoding shown in status bar

Figure 1 shows a screenshot of the Mosaic for Windows Home Page.

Figure 1: NCSA Mosaic.
Source: http://medialab.di.unipi.it/doc/SEHTML2D/figs/02fig09.gif

As can be seen from the above list, Mosaic already had basic browser functionality and features that we have gotten used to, and it worked in a way we are still using browsers today: the client/server principle applied to the Internet.

The Client/Server Principle

The client/server principle is based on a pretty simple idea, illustrated in Figure 2: Interactions between software systems are broken down into two roles: *Clients* are requesting services, *servers* are providing services. When a client wants a service such as a database access or a print function to be executed on its behalf, it sends a corresponding *request* to the respective server. The server will then process this request, i.e., execute the access or the printing, and will eventually send a *reply* back to the client.

This simple scheme, described in more detail, for example, in [26], has become extremely successful in software applications, and it is this scheme that interactions between a browser and a Web server are based upon. A common feature of this principle is that it often operates in a *synchronous* fashion: While a server is responding to the request of a client, the client will typically sit idle and wait for the reply; only when the reply has arrived, the client will continue whatever it was doing before sending off the request. This form of interaction is often necessary; for example, if the client is executing a part of a workflow which needs data from a remote database, this part cannot be completed before that data has arrived. It has also been common in the context of the Web until recently; we will have to elaborate more on this later.

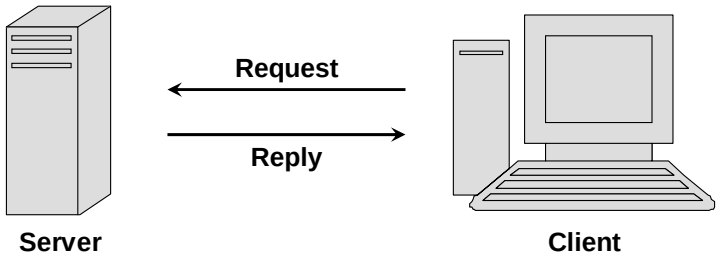

Figure 2: The Client/Server Principle.

We mention that, in a larger network, clients may need help in finding servers that can provide a particular service; a similar situation occurs after the initial set-up of a network or when a new client gets connected. Without going into details, this problem has been solved in a variety of ways; for example, there could be a directory service in which clients can look up services, or there might be a “broker” to which a client has to talk first and which will provide the “address” of a server. Another option, also useful for a number of other issues arising in a computer network (e.g., routing, congestion control, consistent transaction termination), is to designate a central site as the network monitor; this site would then have all the knowledge needed in the network. An obvious drawback is that the network can hardly continue to function when the central site is down, a disadvantage avoided by peer-to-peer networks discussed below.

HTML and HTTP

The basics that led to launching the Web as a service sitting atop the Internet were two quickly emerging standards: HTML, the *Hypertext Markup Language*, and HTTP, the *Hypertext Transfer Protocol*. The former is a language for describing Web pages, i.e., documents a Web server will store and a browser will render. The latter is a protocol for getting a request for a page from a client to a Web server and for getting the requested page in a reply back to the browser. Thus, the client/server principle is also fundamental for the interactions happening on the Web between browsers and Web servers, and, as we will see, this picture has only slightly changed since the arrival of Web *services*. Over the years, HTML has become very successful as a tool that basically everybody can employ to put information on the Web, the reasons for this including the fact that HTML is a vastly fault-tolerant language, where programming errors are simply ignored, and the availability of numerous tools, from simple text editor to sophisticated WYSIWYG (what you see is what you get) environments, for writing HTML documents.

Netscape & Co.

The initial version of Mosaic was launched in March 1993, its final version in November the same year, and although far from modern browser functionality with all their plug-ins and extensions (just take a look at Version 2 of the Mozilla Firefox browser published in the fall of 2006), users pretty soon started to recognize that there was a new animal out there with which one could easily reach for information that was stored in remote places. A number of other browsers followed, in particular *Netscape Navigator* (later renamed *Communicator*, then

renamed back to just *Netscape*) in October 1994 and *Microsoft Internet Explorer* in August 1995; these two soon got into what is nowadays known as the *browser war*, which between the two was won by Microsoft but which is nowadays continuing between Microsoft's Internet Explorer and Mozilla Firefox.

In mid-1994, Silicon Graphics founder Jim Clark started to collaborate with Marc Andreessen to found *Mosaic Communications* (later renamed to *Netscape Communications*). Andreessen had just graduated from the University of Illinois, where he had been the leader of the Mosaic project. They both saw the great potential for Web browsing software, and from the beginning Netscape was a big success (with more than 80% market share at times), in particular since the software was free for non-commercial use and came with attractive licensing schemes for other uses. Netscape's success was also due to the fact that it introduced a number of innovative features over the years, among them the on-the-fly displaying of Web pages while they were still being loaded; in other words, text and images started appearing on the screen already during a download. Earlier browsers did not display a page until everything that was included had been loaded, which had the effect that users might have to stare at an empty page for several minutes and which caused people to speak of the "World-Wide Wait." With Netscape, however, a user could begin reading a page even before its entire contents was available, which greatly enhanced the acceptance of this new medium. Netscape also introduced other new features including cookies, frames, and later JavaScript, some of which eventually became open standards of the W3C and ECMA, the European Computer Manufacturers Association. Figure 3 shows Version 4 of the Netscape browser, pointed to the Netscape homepage of April 1999; it also explains the main features to be found in Netscape, such as the menu bar, the navigation, address, and personal toolbars, the status bar, or the component bar.

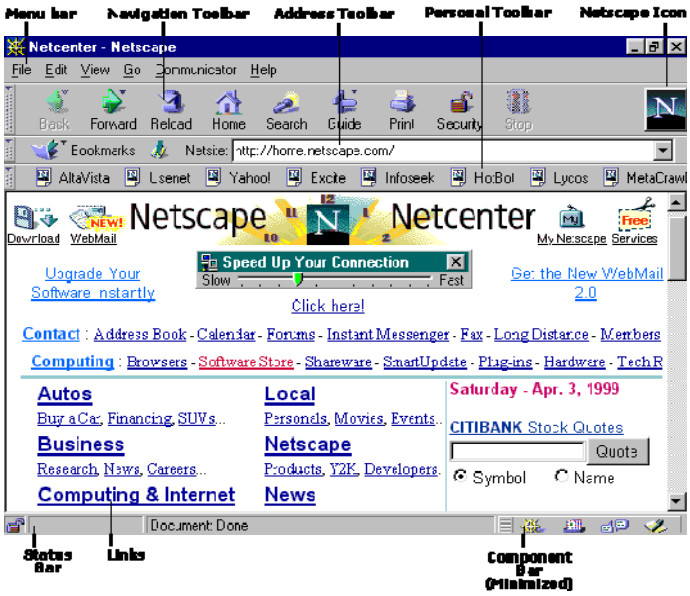

Figure 3: Netscape Navigator 4.
Source: http://www.3dmultimedia.com/help/internet/gifs/navigator_eng.gif

Although free as a product for private use, Netscape's success was big enough to encourage Clark and Andreessen to take Netscape Communications public in August 1995. As Dan

Gillmor writes in his blog at bayosphere.com/blog/dangillmor: “I remember the day well. Everyone was agog at the way the stock price soared. I mean, this was a company with scant revenues and no hint of profits. That became a familiar concept as the decade progressed. The Netscape IPO was, for practical purposes, the Big Bang of the Internet stock bubble – or, to use a different metaphor, the launching pad for the outrages and excesses of the late 1990s and their fallout. … Netscape exemplified everything about the era. It launched with hardly any revenues, though it did start showing serious revenues and had genuine prospects (until the company made critical mistakes and Microsoft decided to abuse its monopoly).”

2.2 The Flattening of the World

The initial public offering (IPO) of Netscape in 1995 is one of the events that Thomas L. Friedman, three-time Pulitzer Prize winner and foreign affairs columnist for the New York Times, in his book *The World Is Flat – A Brief History of the Twenty-First Century* calls a world “flattener” [10]: The world as of today has become a flat one in which people from opposite ends of the planet can interact, play, do business with each other, and collaborate, even without knowing each other, and where companies can pursue their business in any part of the world depending on what suits their goals and intentions best; they can also look at an entire world of customer base. There are essentially no more serious limits to what anyone can accomplish in the world these days, since the infrastructure we can rely upon and the organizational frameworks within which we can move allow for so many unconventional and innovative ways of communicating, working together, collaborating, and information exchange. In total, Friedman accounts for 10 flatteners which in the order he discusses them in [10] are:

1. The fall of the Berlin wall on November 9, 1989, when Eastern Europe opens up as a new market, as a huge resource for a cheap, yet generally well-educated work force, and as an area with enormous demand for investments and renovation. The globalization wave swaps from the West across Eastern Europe and extends deeply into Asia. Millions of people in previously communist countries obtain a new political and economic freedom at a scale they have never known before.

2. The aforementioned Netscape IPO on August 9, 1995, when for the first time in history it was demonstrated that one can make money through the Internet, in particular with a company whose business model does not immediately imply major revenues. Netscape and the Internet become part of a digital revolution immediately picked up by the Eastern countries.

3. Workflow software, which can connect people all over the world by chaining together what they are doing into a comprehensive whole. This software can now even run over the Internet and enable new forms of collaboration and distribution of work. Jobs become location- and time-independent. Indeed, US tax declarations done in Bangalore, India or parts for a Boeing 777 plane manufactured in Russia are becoming normality. A division of labour in specialized tasks has moved from a regional or domestic to an international scale.

4. Open sourcing, or the idea of self-organizing collaborative communities which in particular are capable of running large software projects. The most prominent examples in use today are the GNU/Linux operating system or the Mozilla Firefox browser project. In both cases,

complex software with numerous components has been developed in a huge common effort, and is being maintained by a community of developers which respond to bugs and failures with an efficiency unknown to (and vastly impossible for) commercial software companies.

5. Out-sourcing, where companies concentrate on their core business and leave the rest to others who can do it cheaper and often more efficiently. Out-sourcing occurs at a global scale, i.e., is not restricted to national boundaries or regional constraints anymore. This particularly applies to software production, parts of which are often given to Indian companies or, more recently, also to programmers in Eastern Europe.

6. Off-shoring, which means going way beyond outsourcing; indeed, the next step is to take entire production lines to an area of the world where labour is cheaper. This particularly refers to China since it has joined the World Trade Organization in 2001.

7. Supply-chaining or the idea of streamlining supply and production processes on a world-wide basis, for example through the introduction of RFID tags.

8. In-sourcing, which is the opposite of outsourcing. It sometimes makes sense to bring specific functions into a company in order to have them executed more efficiently. An example is UPS and their handling of Toshiba laptop repairs: UPS not only picks up laptops that need to be repaired from the customers, but nowadays even repairs them in a UPS customer center, in order to be able to ship it back to the customer as soon as the repair has been completed. Apparently, this is way more efficient than shipping a laptop back to Toshiba in Japan, and transporting it from there back to the customers once finished.

9. In-forming thanks to search engines such as AltaVista, Google, Yahoo!, or MSN Web Search. In the flat world, knowledge and entertainment can be obtained at any time and anywhere. Information is accessed through search engines, emails are read on the move, and movies are downloaded on demand. A 21st century person does thus no longer depend on printed newspapers, physical office space, or the local library.

10. Finally, the steroids, i.e., the technological developments which have made everything digital, mobile, personal, virtual, such as modern PDAs, cell phones, data servers as a commodity, cheap personal and laptop computers with high computational capabilities, huge storage capacities, and excellent input/output facilities.

Interestingly, Flatteners No. 5 and 6 are also the subject of a recent ACM report on the globalization and off-shoring of software [2], which contains the findings of a task force that looked into the rapid globalization of IT and the migration of jobs resulting from outsourcing and off-shoring, and which vastly agrees with what Friedman is describing; some findings can even be taken as an extension of Friedman's arguments in the specific direction of the IT industry.

Obviously, not all flatteners are related to the Internet and the Web, yet all of these developments, which go together and influence each other, heavily rely on efficient communication networks and on tools such as the Web for utilizing them. While we are not discussing all the impacts these flatteners have had since the mid-1990s or, as Friedman claims, especially in the (still pretty young) 21st century, some are relevant for us as well: We

have already mentioned Flattener No. 2, the Netscape IPO, which from today's perspective can be seen as one of the initiators of the "dot-com bubble" that erupted during the late 1990s (and that is basically repeating itself today; just take a closer look at the daily trade news); indeed, all of a sudden an Internet company without any highly valued "product" could exhibit an enormous market value. Another is Flattener No. 9, which refers to the entirely changed way in which people handled the Internet, once the Web had arrived. All of a sudden, it became possible to access arbitrarily remote information in an easy and vastly intuitive way ("the global village"), in particular information of which it had not been known before that it existed. One of the slogans now was to have "information at your fingertips," and *search engines* were the tool that made this flattener possible.

2.3 From Linking to Searching

Links and Navigation

The key to what has made the Web so popular early on is the fact that a Web page or an HTML document can contain *hyperlinks* or *links* for short, which are references to other pages (or other places in a current page). The origin of this is *hypertext*, an approach to overcome the linearity of traditional text that was originally suggested by Vannevar Bush in an essay entitled *As We May Think* which appeared in *The Atlantic Monthly* in July 1945 (see http://www.theatlantic.com/doc/194507/bush). Selecting a link that appears in a given HTML document causes the browser to send off a request for the page whose address is included in the link (or, if the link points to another place in the current page, to go that position); this page will then be displayed next.

Figure 4 gives a rough idea of what that means at a larger scale; as can be seen, the Web is a large collection of hyperlinked documents and can be perceived, from a more technical point of view, as a directed *graph* in which the individual pages or HTML documents are the nodes, and in which links leading from one page to another (or back to the same page) are the (directed) edges. Clearly, Figure 4 shows a finite subset of nodes and links only, and can easily be extended in any direction and by any number of further nodes and edges.

Links in HTML are technically *anchors* which typically are composed of a name (that will show up in the document where the links is placed) and a URL, a *Universal Resource Locator* or address of a Web page. When a user clicks on the link, the browser will contact the Web server behind that URL (through common network protocols which, among other things, will ensure name resolution, i.e., translate the URL into an IP address through various steps of address translation) and request the respective HTML document. Links allow a form of *navigation* through the Web, the idea being that if something that a user is looking for is not contained in the current page, the page might contain a link to be followed for getting her or him to the next page, which may in turn be more relevant to the subject in question, or may contain another link to be followed, and so on. Links, however, need not necessarily point to other pages, but can also be used to jump back and forth within a single page or they can link to different types of content (e.g., images, videos). The reader interested in a technical analysis of navigation should consult [13].

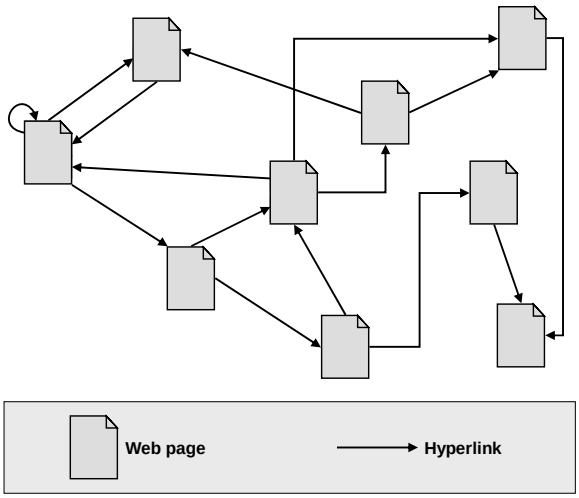

Figure 4: Navigation through the Web along hyperlinks.

From a somewhat conceptual point of view, it is fair to say that the Web is a *very large* graph (with a current – fall of 2006 – estimate of more than 10 billion nodes) and an even larger number of edges. Following links then means creating *paths* in that graph, i.e., sequences of nodes that have a source and a target. However, paths are not formed randomly, but typically represent a *search* for a term or a combination of terms contained in a node (Web page). The size of the Web then makes it impossible to search comprehensively, since there is hardly any *a priori* indication of how long a search would have to continue, or how long a search path could become. This situation is complicated by the fact that the Web as a graph is never even known completely in advance. In fact, the Web is a *dynamic* graph in which both nodes and edges come and go. Moreover, parts of the Web might be unreachable at a time due to network problems, or Web designers may add new pages with links and from time to time remove old ones. Investigations of the connectivity of the Web as a graph have revealed that it is not as interconnected as one might think. Indeed, as reported in *Nature* 405, 113 (May 2000), the Web has the form of a “bowtie” with several components: “A central core contains pages between which users can surf easily. Another large cluster, labeled 'in', contains pages that link to the core but cannot be reached from it. These are often new pages that have not yet been linked to. A separate 'out' cluster consists of pages that can be reached from the core but do not link to it, such as corporate websites containing only internal links. Other groups of pages, called 'tendrils' and 'tubes', connect to either the in or out clusters, or both, but not to the core, whereas some pages are completely unconnected.” As a consequence, a user might not be able to find a click path between two randomly chosen Web pages, for the simple reason that they are not connected!

Web Exploration

As a consequence of the above, in particular the size of the Web and the fact that it is highly dynamic, classical graph techniques are hardly applicable to the Web, since, for example, there is no way to establish an adjacency matrix for the Web, a typical prerequisite for many algorithms on graphs. Instead of traditional traversals through a graph, the Web relies more on

exploration, i.e., the progression along paths or sequences of nodes without predetermined targets.

Exploration has an extremely practical perspective, namely the question of how to find specific information in this "hyperspace," or how to make sense to this large and dynamic graph for a particular application or usage at hand. This is where the activity of *search* comes in. In the early days of the Web, exploration and search were soon identified as being relevant, if not even the major form of getting to information on the Web; however, automatic tools were not yet around. This was the time when a new type of job was created, the *information broker*. In the mid-1990s, clients could turn to an information broker for having a search done on the Internet (and in addition through more traditional information sources). If the client was a company, which was most often the case, the broker would typically come back with market data, data on market development, information on the competition both domestically and internationally, typical figures of how much is spent on advertising and what common distribution channels are, and real or estimated consumer demand. Clearly, that not only required an in-depth investigation, but also analytical capabilities, i.e., ways to distinguish highly relevant information from less relevant and to aggregate the former into concise, readable and often graphical statements, and often inside knowledge of the particular area the customer was from. Indeed, there is a difference whether an information broker works for a pharmaceutical company or for a construction company that is building high-speed trains or exclusive ocean liners.

While the information broker as a job description has lost importance over the years thanks to the automated tools for doing this job we will discuss next, an important aspect is still around today, that of *price comparisons*. Indeed, comparing prices over the Web has become an important activity, for both companies and individual users, and is a form of information brokering still available today (through sites such as www.dealtime.com, www.mysimon.com, www.bizrate.com, www.pricewatch.com, or www.pricegrabber.com, to name just a few).

Web Search

Search *engines* are today's most important tool for finding information on the Web, and they have emerged relatively soon after the Web had been launched in 1993. Although "to search" the Web is nowadays often identified with "to google" the Web with Google getting roughly 45% of all the search traffic today (searchenginewatch.com/2156431), Google has not been the first search engine around, and will most likely not be the last. However, it has dominated the search field ever since its launch in the fall of 1998, and it has invented many tools and services now taken for granted; the interested reader should consult [6] for the original research on Google, [27] for an account of the history of Google, and [15] for an in-depth presentation of its possibilities. [13] describes how search engines work in general. For fairness reasons, we mention that AltaVista, Yahoo!, InfoSeek, Inktomi, AlltheWeb, Ask, Vivisimo, A9, Wisenut, or Windows Live Search are among the many other search engines out there.

Search has indeed become ubiquitous. People nowadays search from the interface of a search engine, and then browse through an initial portion of the often thousands or even millions of answers the engines brings back. Search often even replaces entering a precise URL into a browser. In fact, search has become so universal that John Battelle in his book *The Search* speaks of the *Database of Intentions* that nowadays exists on the Web [4]: It is not a materialized database stored on a particular server, but "the aggregate results of every search

ever entered, every result list ever tendered, and every path taken as a result." He continues to state that the Database of Intentions "represents a real-time history of post-Web culture – a massive click stream database of desires, needs, wants, and preferences that can be discovered, subpoenaed, archived, tracked, and exploited for all sorts of ends." Search not only happens explicitly, by referring to a search engine; it also happens to a large extent inside other sites, for example within a shopping or an auction site where the user is looking for a particular category or product; also most newspaper sites provide a search function that can be used on their archive. As a result, a major portion of the time nowadays spent on the Web is actually spent searching, and Battelle keeps watching the developments in this field in his search blog on the topic (http://battellemedia.com/).

From a technical perspective, a search engine is typically based on techniques from information retrieval (IR) as explained, for example, in [13] or [23] and has three major components as indicated in Figure 5: A crawler, an indexer, and a runtime system. The *crawler* explores the Web as indicated above and constantly copies pages from the Web and delivers them to the search engine provider for analysis. Analysis is done by an *indexer* which extracts terms from the page using IR techniques and inserts them into a database (the actual *index*). Each term is associated with the document (and its URL) from which it was extracted. Finally, there is the *runtime system* that answers user queries. When a user initiates a search for a particular term, the indexer will return a number of pages that may be relevant. These are ranked by the runtime system, where the idea almost always is to show "most relevant" documents first, whatever the definition of *relevance* is. The pages are then returned to the user in that order.

The popularity of Google grew out of the fact that they developed an entirely new approach to search. Before Google, it was essential to locate just any site whose content was related or contained a given search term. To this end, search engine builders constructed indexes of Web pages and often just stored the respective URLs. As an answer to a query, a user would get back a list of URLs through which he or she then had to work through. Google co-founder Larry Page came up with the idea that no all search results could be equally relevant to a given query, but unlike the information broker, who can exploit his or her expertise on a particular field, an automated search engine needs additional ways to evaluate results. What Page suggested was to rank search results, and he developed a particular algorithm for doing so; the result of that algorithm applied to a given page is the *PageRank*, named after his inventor. The PageRank of a page is calculated using a recursive formula (see http://infolab.stanford.edu/~backrub/google.html for details) which we are not discussing here, but the underlying idea is simple. Monika Henzinger, Google research director, explained it in an interview with the German edition of MIT's *Technology Review* in April 2004 using the following analogy: Consider a doctor. The more people recommend the doctor, the better he or she is supposed to be. It is similar with ranking a Web page: The more pages link to a page p, the higher the rank of p will be. However, the quality of a doctor also depends on the quality of the recommender. It makes a difference whether a colleague or a salesperson for the pharmaceutical industry recommends her or him. If the doctor is recommended by another doctor, that recommendation will count 100%; a recommendation from a nurse without comprehensive medical education will count only 60%, that from a patient 20%, and that from the salesperson, having an interest completely disjoint from that of the doctor, will count 0%. The principle behind this, also found, for example, in classical scientific citations, is thus based on the idea of looking at the links going into a page

p in order to calculate the rank of p, but to do so by recursively ranking all pages from which these incoming links emerge. The idea was first explored while Google founders Sergey Brin and Larry Page worked on a project called *BackRub* at Stanford University; over the years, Google has added other criteria for constructing the order in which search results are presented to the user besides PageRank.

The comparison between an information broker, who can apply intuition, human expertise and reasoning, as well as experience and domain knowledge to search results in order to distinguish good and bad ones, and a search engine, which has to do all of this based on some form of artificial intelligence, points to the fact that even beyond Google there is lots of room for improvement in search. For example, a ranking algorithm may learn from the user as a search progresses, an idea behind the Mooter search engine (www.mooter.com). For another example, Yahoo! Mindset is based on the idea, shown in Figure 6 for the search term “bank”, that a user may be somewhere in a “shopping” and a “research” mode or somewhere in between during the search, depending on whether the user is looking for commercial or non-commercial information. Mindset offers a slider (to be seen in the upper left portion of Figure 6, which the user can move left or right in order to get the search results reordered according to a changed mode. By the same token, other novel ideas have been developed around search and realized in engines such as Ask, Wisenut, Vivisimo, TransparenSee, and others. For an up-to-date account, the reader should consult searchenginewatch.com and consult their “Search Engine Resources” section.

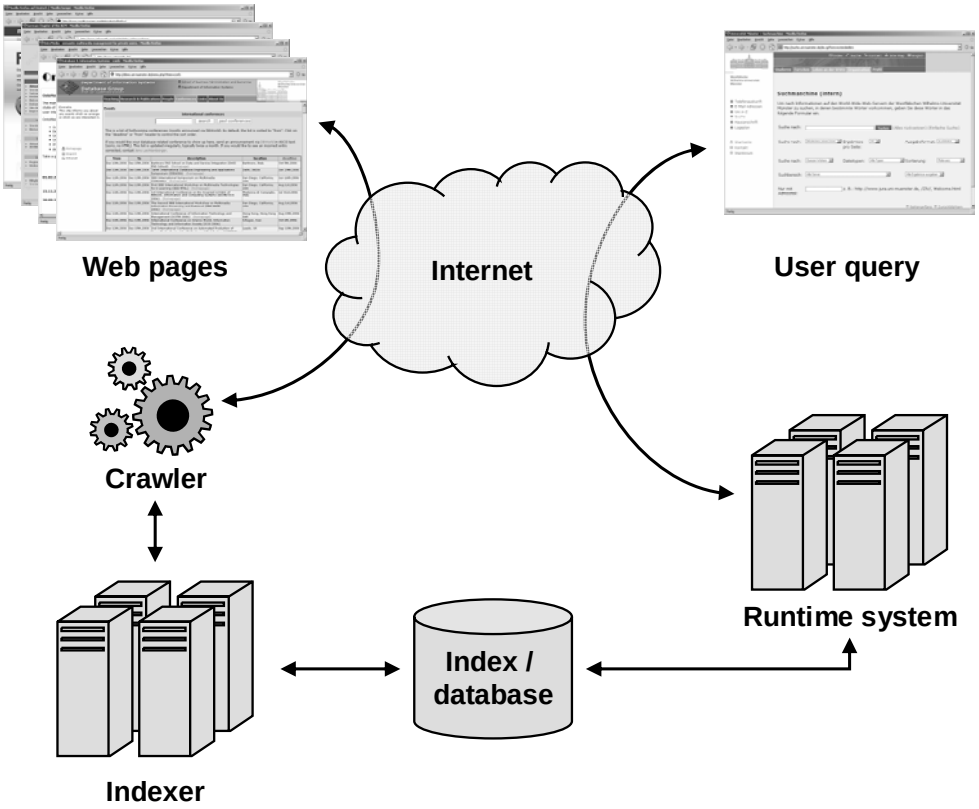

Figure 5: Anatomy of a Search Engine.

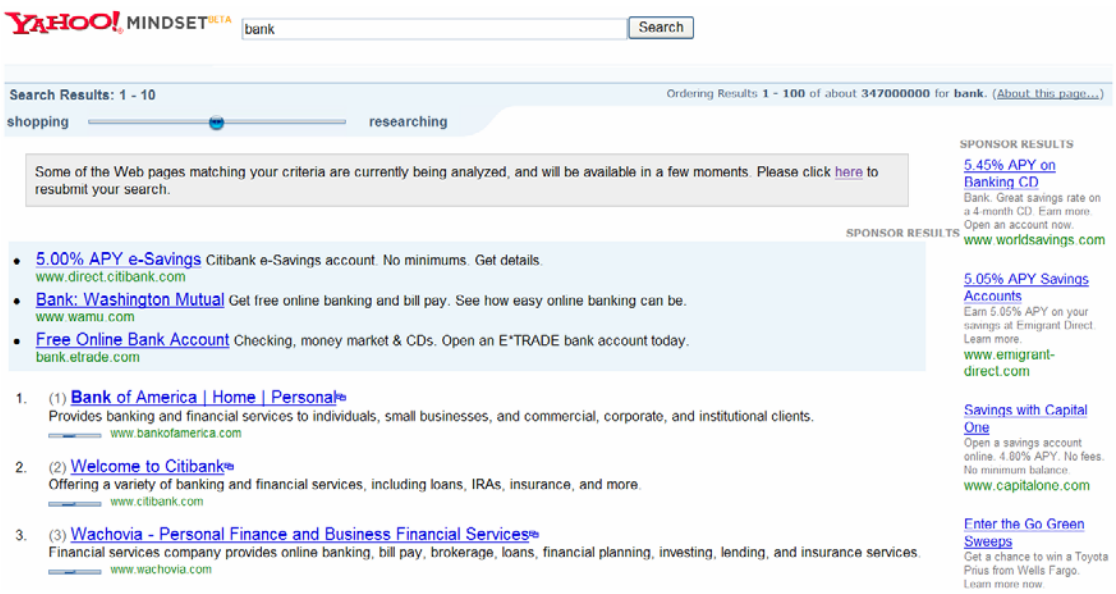

Figure 6: Yahoo! Mindset search result for term “bank.”
Source: http://mindset.research.yahoo.com/search.php?p=bank

The Long Tail of Search

It is interesting to look at some statistics about searching. To the end, readers may consult, for example, Google’s Zeitgeist (at www.google.com/zeitgeist), which keeps rankings about most popular search terms in recent past. For example, the 5 search queries entered at Google in 2006 with the biggest gain compared to 2005 were

- bebo
- myspace
- world cup
- metacafe
- radioblog

(see google.com/intl/en/press/zeitgeist2006.html). Other statistics may be obtained from places like Nielsen Netratings or the aforementioned searchenginewatch.com. What people have observed by looking at these figures is, among other things, that few queries have a very high frequency, i.e., are asked by many people and pretty often, but then a large number of queries have a considerable lower frequency. When plotted as a curve, where the x-axis represents a list of (a fixed number of) queries, while the y-axis indicates their frequency, the graph will look like the one shown in Figure 7. Graphs of this form follow a power-law type of distribution: They exhibit a steep decline after an initial, say, 20%, followed by a massive tail into which the graph flattens out. Power laws can be found in many fields: the aforementioned search term

frequency, book sales, or popularity of Web pages. Traditionally, when resources are limited, e.g., space on a book shelf or time on a TV channel, the tail gets cut off at some point. The term "long tail" is used to describe a situation where such a cutting off does not occur, but the entire tail is preserved. For example, there is no need for search engine providers to disallow search queries that are only used very infrequently. As we will see as we go along in this book, the long tail phenomenon pops up in a variety of contexts related to the Internet and the Web. The long tail is discussed in detail in Chris Anderson's book [3] as well as on his Web site (http://www.thelongtail.com/).

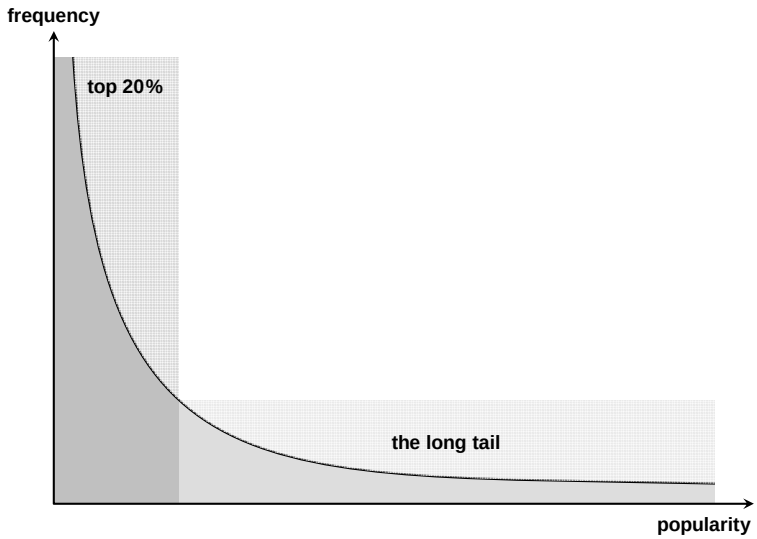

Figure 7: The Long Tail (of Search Queries).

Portals

We mention that Yahoo! and AOL were among the first to recognize that the Web with its exploding number of pages already in the mid-90s needed some form of organization, and they did so by creating *portals* containing categorizations of Web site content and pathways to other content. A typical portal page is shown in Figure 8 with its categories such as *Arts, Games, Health, News, Recreation,* etc. Each such category will typically contain a list of subtopics, eventually together with the current number of hits in each. For example, hitting "Recreation" on dmoz.org will reveal a list whose first 10 entries are as follows (where the "@" sign indicates membership in another category as well):

Antiques (1,050)
Astronomy@ (1,339)
Audio (211)
Autos (8,702)
Aviation (4,842)
Birding (1,948)
Boating (2,863)
Bowling@ (663)
Camps (1,123)
Climbing (1,402)

Portals are actually a bit older than search engines and originally even did not accept the idea that search was a necessary feature for the Web. The reason behind this is easily identified as being commercial: If a portal has sites included in its categories and places banner ads on pages, it will be interested in many people using the portal and its listings, so that the ads can drive home some revenue. But this will only work if traffic is not "distracted" to other sites, which may not be listed in the portal, by a search engine. In other words, portals were originally afraid that search engine take away too much of the traffic that would otherwise reach them.

If a user is looking for very specific information, using a portal may today not be a rewarding experience, since it may not be obvious under which category to look for the term in question. Therefore, it has become common even for a portal to offer a search field which can freely be used for arbitrary searches inside the portal. The modern version of classical portals such as Yahoo! do still provide a categorization, yet this may be more difficult to find (e.g., under dir.yahoo.com), since other services have become predominant. In other words, the original portal idea of providing a single point of access to the Web still exists, yet usability has gained much more priority over the years and has modified the appearance of portals considerably. For example, a quick look at www.yahoo.com in Figure 9 shows that, from the home page, services such as *Mail*, *Messenger* or *Weather* are only one click away.

Portals can not only be seen as an alternative (or competition, depending on the point of view) to search engines, where potential search results have been categorized in advance; they are often also highly specialized, for example towards a particular business branch or interest. Examples for the former include travel booking portals such as Travelocity, Hotels, RatesToGo, or Expedia, one for the latter is the recently launched German informatics portal io-port.net, which "offers fast and convenient access to more than two million publications in informatics and related subject areas from all over the world. All information, which up to then had been stored in various data sources, has been consolidated and is now available from one source. All steps required for information retrieval are available via an easy-to-use, customizable interface and supported by powerful tools." Thus, portals even today have lost little of their popularity, and new portals are still being launched from time to time. An example of such a recent addition to the portal scene is Kosmix (www.kosmix.com), which specialized in fields like health, video games, finance, travel, US politics, and autos.

As we will discuss later, the next evolutionary step is that a user can configure a "home page" entirely according to his or her preferences and services needed. Actually, even the Yahoo! starting page is partially configurable, but sites such Netvibes (www.netvibes.com) take this idea a considerable step further. Here, the user can choose from a rich menu of different search, mail, or blog services, which are then "mashed" together into a personalized page.

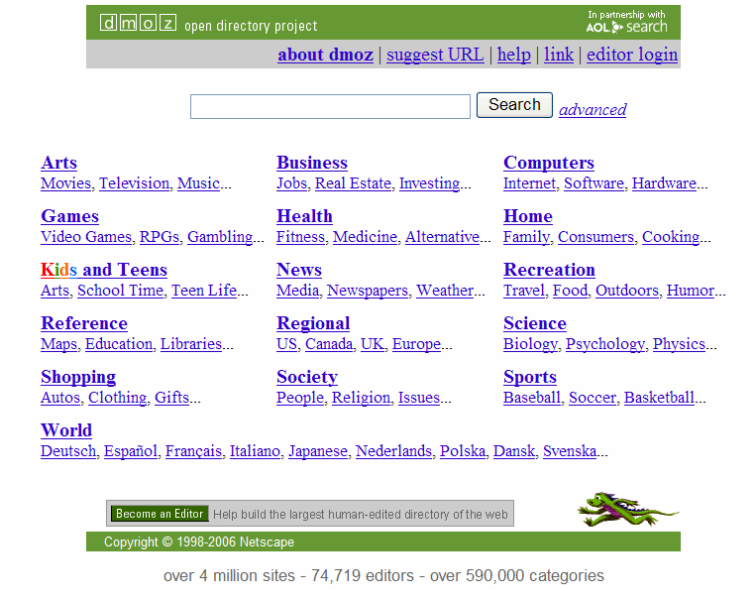

Figure 8: The dmoz.org example of a portal page.
Source: http://dmoz.org/

Figure 9: The Yahoo! home page as of November 2006.
Source: www.yahoo.com

2.4 Commercialization of the Web

Roughly during the mid-90s people started discovering that there is also a commercial side to the Web. We have already mentioned the Netscape IPO, but commercialization was and is not just about buying (and eventually selling) Internet companies. Commercialization has in particular materialized in the form of *electronic commerce*, commonly abbreviated as *e-*

commerce, which involves moving a considerable amount of shopping and retail activity essentially from the street to the Web. More generally, e-commerce refers to selling goods or services over the Internet or over other online systems, where payments may be made online or otherwise. It was typically during the weeks before Christmas in which the success as well as the growth of e-commerce could be measured best every year (see, for example, Bertil Lindberg's collection of Web pages on global e-business and e-commerce at http://home.earthlink.net/~lindberg_b/GEC.htm for a number of statistics on the subject). In the beginning, customers were reluctant to do electronic shopping, since it was uncommon, it was not considered an "experience" as it may well be when strolling through physical shops, and it was considered unreliable. Many companies entering this new form of business were not ready yet, unaware of the process modifications they would have to install in their front- and back-offices, and unfamiliar with the various options they had from the very beginning. Major obstacles in the beginning also were lacking security, in particular when it came to payments over the Web, and lacking trust, in particular when it came to the question of whether goods I had paid for would indeed be delivered to me. As a consequence, e-commerce took off slowly in the very beginning. However, the obstacles were soon overcome, for example by improvement in hardware and software (e.g., session handling), by encrypting payment information or by the introduction of trusted third parties for handling the physical aspects of sales transactions. Then, towards the end of the century, e-commerce started flying, with companies such as CDnow and Amazon.com, later also eBay, and sales figures soon went beyond the billion-dollar threshold. Today, even retail chains such as Walmart or Costco make considerable revenues online, in addition to the numerous stores they run in a traditional way.

However, it was also discovered that e-commerce and selling over the Web was not the only way of making money on or through the Web. Indeed, another was placing advertisements, and ultimately to introduce "paid clicks" (to be discussed below). Besides all this is, of course, the telecommunication industry, for which technological advances such as the arrival of DSL or wireless networks brought entirely new business models for both the professional and the private customer.

CDnow

CDnow is a very typical example of how doing business over the Web took off. CDnow was created in August 1994 by brothers Jason and Matthew Olim. As they describe in their personal account of the company [20], it was started in the basement of their parents' home; Jason became the president and CEO and Matthew the Principal Software Engineer. The company was incorporated in Pennsylvania in 1994 and originally specialized in selling hard-to-find CDs. It went public in February 1998, and after financial difficulties eventually merged with Bertelsmann, the big German media company, in 2000. CDnow became famous for its unique internal music rating and recommendation service, which was also often used by those who had never actually purchased a product on the site. In late 2002, Amazon.com began operating the CDnow web site, but discontinued CDnow's music-profiling section.

What the Olim brothers detected early on was that the Web offered a unique chance to provide not only information, but highly specialized information that previously had required an enormous amount of research to come by. This is what they provided for music on CDs, and they combined their information and catalogue service with the possibility to buy CDs directly from them. At some point, CDnow was probably the best online store for music, as it was able to

integrate so much information on a CD, on an artist, or on a group in one place and in so many distinct categories. Their selection was enormous, and most of the time whatever they offered could be delivered within days. They also ran into problems that nobody had foreseen in the very beginning, for example that customs fees may need to be paid when a package of CDs is delivered to an addressee in a foreign country. In other words, legal issues related to commerce over a network that does not really have physical boundaries, came up in this context (as it did for any other shop that now started selling internationally), and many of these issues remain unresolved even today. This is different for non-physical goods, which can be distributed electronically and for which these issues do not exist. iTunes is currently by far the most popular service for distributing music electronically.

E-Commerce

E-commerce, as said, is nowadays understood as the process of buying, selling, transferring, or exchanging products, services, or information via computer networks and in particular the Web; see Laudon and Traver for a comprehensive introduction [12]. It has become so popular since it provides an innovative approach for conducting business for many companies, often in addition to a traditional business, it represents a reduction in the cost of transactions, it can provide unique, customized products for even small customer bases, and it allows customer access 24 hours a day, 7 days a week ("24/7"). Moreover, e-commerce *between* customers has been highly popularized through auctioning sites. Often, a fundamental characteristic of an e-commerce scenario is the absence of *intermediaries,* i.e., third parties offering intermediation services to two trading parties. For example, a publishing company can now sell directly to readers, without going through the intermediary of a book store, or individuals can sell used cars without the help of a car dealer.

A typical e-commerce system has four major components:

- **Product presentation component:** An e-commerce system must provide a way for customers, which can be individual consumers, companies, or a business, to search, select, and identify products they want to purchase. Many e-commerce web sites have an electronic catalogue which is a list or categorization of products with descriptions, pictures, and prices. Commonly and as mentioned, some form of search function is provided that aids inspecting the catalogue.

- **Order entry and shopping basket component:** After the customer has selected a desired product, he or she needs to enter an order for the product into the electronic commerce system. Order entry often allows the customer to add items to an electronic shopping basket, which is a list of the products the customer wants to purchase. Before an item is added to that basket, the e-commerce system should have the inventory control system check the product database to see if there is adequate stock on hand or if the product needs to be ordered from a manufacturer.

- **Payment component:** To allow customers to pay for the items purchased, an e-commerce system needs to have an electronic payment capability. Various approaches are used for electronic payment, including payment by credit card or by electronic funds transfer. To ensure the security of the card information sent over the Internet, special protocols such as HTTPS that provide data encryption are used.

- **Customer service and support component:** At any time before, during, or after purchasing an item, the customer may need advice or special services. For example, a customer may have a question about how a particular item of clothing fits before purchasing it. During the ordering process, the customer may have difficulty using the electronic commerce system. After receiving an item, the customer may decide to exchange or return the item. Some time later, the customer may have a warranty claim. Many of these customer service situations can be dealt with by providing detailed information and answers to questions electronically. A well-designed e-commerce system will therefore provide capabilities for customer service.

Clearly, while these components are well-understood these days, it has taken more than 10 years of development and experience to come up with this understanding. In the beginning, user acceptance of e-commerce was low, due to limitations in Internet access, to the limited number of companies doing e-business at all, to the mentioned lack of trust, or due to the missing customer support. Indeed, a traditional retail shop will know returning customers after a short while, whereas an e-commerce site cannot connect an HTTP request by a customer today from one by the same customer tomorrow. It also took companies a while to recognize that doing their business electronically involves much more than setting up a Web site that has the components listed above. Indeed, it requires a considerable amount of process reengineering, as, for example, the Web shop and the back office must be connected in novel ways that did not exist before; this has been overlooked frequently in the early days of electronic commerce and was a typical reason for failure. Apparently among the fastest to realize this and to react properly have been banks and financial institutions, since their business has moved to the Web quite considerably; electronic banking stock trade as well as fond and portfolio management are in wide use today.

Meanwhile, e-commerce has triggered the development of a whole new branch of the software industry, which not only provides software to install and run a Web shop, but also systems for click stream analysis, data mining [9], and customer relationship management (CRM) [21]. Many software systems for e-commerce nowadays are *agent-based*, and comprise programs called *agents* for tasks such as the identification of customer needs, product search, finding best bargains, price negotiations, arrangement of payment or delivery, after-sales services, advertisement, or fraud detection. As an example of an e-commerce architecture for an enterprise, Figure 10 shows a typical set-up of the Enfinity Suite 6 from the German software company Intershop.

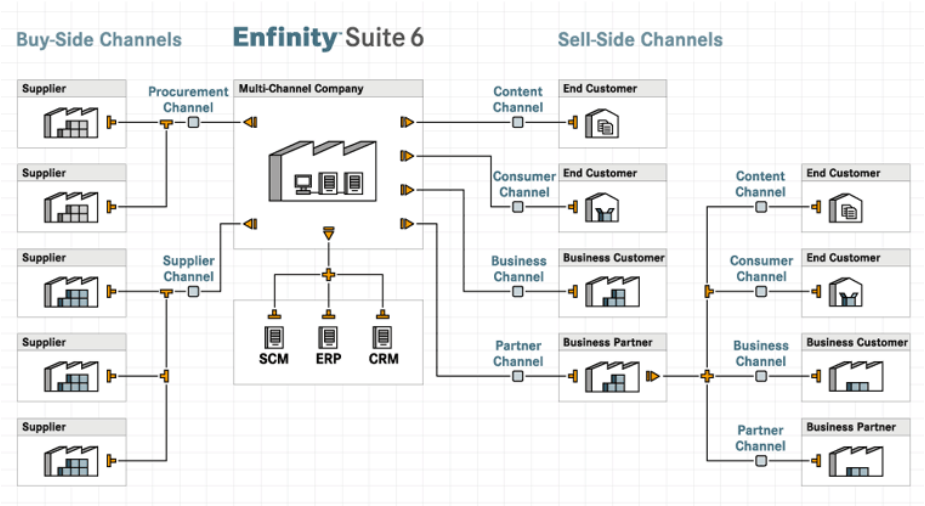

Figure 10: Intershop Enfinity Suite 6.
Source: http://www2.intershop.com

As can be seen here, an e-commerce platform nowadays serves both the buyer as well as the seller side. On the buyer side, there are typically a number of suppliers from which the company is getting its raw materials or supplies, often through some form of procurement process, and, thanks to the flattening we have mentioned, this process can be executed world-wide. Internally, there is a supply-chain management (SCM) system at work that interacts with an enterprise resource planning (ERP) system in various ways. On the seller side, there may be several channels through which the company sells its goods or services also world-wide, including consumers, businesses, and partners. For their various customers, some form of CRM-system will be in place in order to take care of after-sales activities, customer contacts, complaints, warranty claims, help desk inquiries, etc.

Forms of E-Commerce

We mention that electronic commerce has developed into various forms that all have their specific properties and requirements:

- Business-to-business or B2B e-commerce, which involves a business selling its goods or services electronically to other businesses.
- Business-to-customer or B2C e-commerce, which involves a business selling its goods or services electronically to end-customers.
- Customer-to-customer or C2C commerce, which involves a customer selling goods or a service electronically to other customers.

B2B e-commerce, in turn, comes in three major varieties: In a *supplier-oriented marketplace*, a supplier provides e-commerce capabilities for other businesses to order its products; the other businesses place orders electronically from the supplier, much in the same way that consumers will place orders in B2C e-commerce. In a *buyer-oriented marketplace*, the business that wants to purchase a product requests quotations or bids from other companies electronically; each supplier that is interested places a bid electronically and the buyer selects the winning supplier

from the submitted bids. Finally, in an *intermediary-oriented marketplace*, a third-party business acts as an intermediary between the supplier and the buyer; the intermediary provides e-commerce capabilities for both suppliers and buyers in order to identify each other and to electronically transact business.

B2C e-commerce is probably best known to the general public, although by figures alone B2B is considerably higher in value of goods traded. B2C e-commerce has become popular due to two aspects: the end of intermediaries and better price transparency. Indeed, goods are now often sold directly by a business to the end-customer, instead of going through a third-party in between. For example, software can now be downloaded from the producer directly (called *electronic software distribution*), instead of having it burnt on a CD-ROM or DVD and sold through stores. As a result, prices may be lower (an expectation often not valid), or the seller will make a better profit (since nothing needs to be paid to the intermediary). Second and as mentioned, it has become very popular on the Web to provide price comparisons, through sites such as dealtime.com or guenstiger.de; as a consequence, the customer will nowadays often do extensive comparisons before committing to a particular seller.

The third major type of e-commerce, C2C, is mostly manifested these days through auctions such as eBay.com or TradeMe.co.nz. eBay was founded in September 1995 by computer programmer Pierre Omidyar under the name AuctionWeb. One of the early items sold on eBay was Omidyar's broken laser pointer, which to his surprise was due to a real interest in such an item by the winning bidder. The company officially changed the name of its service from AuctionWeb to eBay in September 1997. Millions of collectibles, appliances, computers, furniture, CDs, DVDs, musical instruments, diecast models, outdoor equipment, cars, and other items are listed, bought, and sold daily on eBay. Some items are rare and valuable, while many other items would have discarded if eBay with its thousands of bidders worldwide would not exist. Anything used or new can be sold as long as it is not illegal or does not violate the eBay *Prohibited and Restricted Items policy*. Interestingly, programmers can create applications that integrate with eBay through the eBay application programming interface (API) by joining the eBay Developers Program. This opens the door to “mashing up” eBay with totally different applications. We mention that other forms of e-commerce have meanwhile emerged, including G2C (government-to-customer) or B2G (business-to-government), to name just two.

Yet another form of commercialization of the Web is what has been happening in the late 1990s (roughly between 1997 and 2001) in the context of the so called *dot-com bubble*, namely the buying and selling of entire Internet companies (the “dot-coms”). With the story of Netscape, we have already mentioned a very prominent example. In 1998 Netscape was taken over by AOL. There are quite a few other examples of start-ups that had a new idea for an Internet site or service in the first place, were taken public pretty soon after the incorporation, became subject to speculation, and were finally bought by another company; in the course of these happenings, many were heavily overvalued and hence overpriced, what together with rapidly increasing stock prices and the temptation to many individuals of being able to make money on the stock market fast ultimately led to the famous crash. Clearly, a number of companies survived the burst of the bubble, among them Google.

Customer Feedback and Recommendations

For anyone interested in setting up an e-commerce site, two questions will prevail: What is an appropriate way to sell my product, and how do I attract traffic to my site? Regarding the former, an obvious difference exists between electronic goods, such as software or music files, and physical goods, such as cars or clothing. Electronic goods are basically easy to sell: After having made the payment, the customer is given an access code and a URL from where he or she can download what was just paid for; the files just purchased will be made available at that address, and will typically be kept there for some period of time in case the download fails and has to be retried. The same is no longer true for non-electronic goods. Indeed, the latter might require that the customer needs to be given advice or data for comparison (in the case of cars) or needs to be given the opportunity to return the merchandise if it does not fit or meet the expectations (an issue nowadays legally regulated in many countries).

Especially for non-electronic goods, it has become very helpful if such *advice comes from other customers*. On places like Amazon.com, eBay and others, this has become one of the main aspects people watch out for when shopping for a product: what others have said about that product before, how they like it, whether or not they would buy it again, and maybe even how its seller has been rated or the overall experience has been. Once the importance of other customers' opinions, evaluations, and recommendations had been recognized, many Web shop providers started to install possibilities for commenting on a regular and intensive basis. Amazon.com was among the first to take this even a step further and go from pure reviews, which can even be commented by others, to a comprehensive recommendation system ("Customers who bought this item also bought …"), whose goal is, as an Amazon employee once put it, "to make people buy stuff they did not know they wanted."

While reviews often come from a professional source (such as the publisher of a book or newspaper staff) or from private customers, recommendations are generated by the data mining tools that work behind the scenes of the Amazon Web site. Recommendation systems look primarily at transactional data that is collected about each and every sales transaction, but also at previous user input (such as ratings) or click paths, and then try to classify a customer's preferences and to build a "profile"; recommendations will then be made on the basis of some form of "similarity" of items or categories that have been identified in consumers' profiles. Clearly, recommendations point to other items, where more customer reviews as well as further recommendations to more products can be found. More recently, Amazon.com has even begun to allow registered users to place *tags* on items, which are explained as follows: "Think of a tag as a keyword or category label. Tags can both help you find items on the Amazon site as well as provide an easy way for you to 'remember' and classify items for later recall." The activity of placing a tag is called *tagging* and has become very popular in recent years.

Notice that the possibility for customers and others to place reviews or tags on an e-commerce site marks the arrival of "user input" to the Web, an option which is included in many serious commerce sites today. Indeed, ratings and comments have been shown to have a major impact on revenues a seller may be able to obtain, and that is no surprise: If a seller is getting bad ratings repeatedly, why would anyone buy from him in the future? This input is typically exploited in various ways, including the offer of a seemingly customer-specific "Gold Box" at Amazon.com or the formation of "*virtual communities*." Such communities are characterized by the fact that its members might not know each other, but they all share common interests. This

phenomenon has been identified and studied by many researchers in recent years, and it represents a major aspect of the socialization of the Internet and the Web.

In the context of electronic commerce, user input to the Web typically comes in one of several forms: It may consist of collecting feedback, comment, reviews or evaluations. Beyond e-commerce, however, it may consist of having a discussion on a specific topic or just producing a monologue of your own opinion or experience in diary form, activities that have become known as *blogging*; we will say more about this topic later. As we will see, providing user input to the Web, or being able to "write" to the Web, is one of the major characteristics of the Web today.

The feature of providing feedback and ratings to an e-commerce site has only been a first step in the direction of turning a shop more and more into an application or even a collection of applications. A typical example is the *wish list* functionality of Amazon.com, where a user can save search results for future purchase and gifts (and delete items that he or she no longer wishes to get). Another such example is the aforementioned tagging functionality, through which users can create their personal "subdatabase" out of the Amazon.com database. Using tags, he or she can structure or categorize items according to personal taste, needs, or demands. Apparently, functionalities like these go far beyond the nice book store as which Amazon.com started selling over the Web back in 1995. Today, the goal is to make a visit of this site a "rich" experience, similar to visiting a (book) store downtown or in the mall, so that the user will hopefully use the site for more and more of his or her shopping needs.

Advertising and its Long Tail

The other aspect the owner of an e-commerce business will be interested in, the attraction of traffic as already mentioned above in the context of portals, is closely related to classical disciplines from business administration: advertising and marketing. Traditionally, *advertising* is the art of drawing public attention to goods or services by promoting a business, and is performed through a variety of media. It is a mechanism of *marketing,* which is concerned with the alignment of cooperations with the needs of the business market. Advertising has become a big business on the Web. It has started out with banners that could be placed on other Web sites. It has meanwhile emerged into one of the major ways to make money on the Web, which according to [4] is due to Bill Gross and his invention of goto.com, a site that does not exist anymore, but has become famous for being among the first to differentiate traffic. Indeed, what Gross, who had founded IdeaLab (http://www.idealab.com/) for helping start-up companies to establish their business, figured was that advertisement on the Web is vastly irrelevant and cannot be beneficial as long as the traffic passing by any placed ad is the "wrong" traffic, i.e., coming from people not interested in what is being advertised. If people arrive at a site due to a spammer who has led them there, due to a bad portal classification, or due to a bad search result, they are hardly interested in the products or services offered at that site. He hence started investigating the question of how to get qualified traffic to a site, i.e., traffic with a basic intention to respond to the goods or services found at a site, and then started calculating what businesses might be ready to pay for this. This gave birth to the idea that advertisement can be associated with the terms people search for and pay-per-click tracking models we see today in this business. Companies like Accipiter (http://www.accipiter.com) or its new mother aQuantive (http://www.aquantive.com/) have even made this part of the software business.

Advertising has become a major business model in particular since the arrival of Google AdSense, “a fast and easy way for website publishers of all sizes to display relevant Google ads on their website's content pages and earn money” and Google AdWords, which allows businesses to “create ads and choose keywords, which are words and phrases related to [their] business. … When people search on Google using one of [the] keywords, [the] ad may appear next to the search results.” Again, we delay a more detailed discussion of advertising as a general business model, and AdSense/ AdWords in particular a little bit. An important point to notice here is that through advertising on the Web, another copy of the “long tail” curve of Web applications we have seen in Figure 7 materializes: Through Google AdWords, it has become possible not only for large companies (amounting to 20% of all companies) to place advertisements on the Web, but now the same can be done even for a small company. Through a cost-effective and highly scalable automated infrastructure provided by the Google search index, Google offers advertising even for very limited budgets as they may be only available for a small company. In other words, small companies do not have to set up an infrastructure for advertising (potentially even in niche markets) themselves, but they can simply rely on what others are providing and searching for on the Web.

Trusted Third Parties

While we have mentioned that, from a seller’s perspective, doing business over the Web may be attractive due to the absence of intermediaries which may take a little share off the profit, there are situations in which new intermediaries enter the picture. This in particular refers to making payments, for which *trusted third parties* have come onto the scene. The term is borrowed from cryptography, where it denotes an entity enabling interactions between two parties who both trust the third party; so that they can utilize this trust to secure their business interactions. Figure 11 shows a typical scenario, here using X-CART, another provider of e-commerce store software, and PayPal as an example.

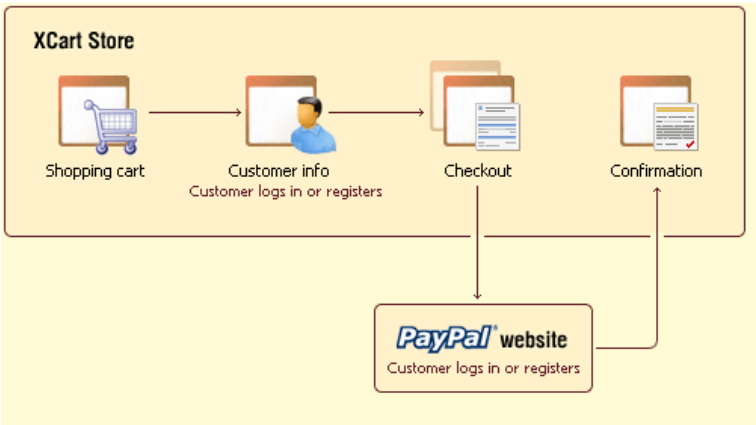

Figure 11: Secure payments through a third party.
Source: http://www.x-cart.com/images/paypal.gif

PayPal allows payments and money transfers to be made over the Web, actually to anybody with an email address, and it nowadays performs payment processing for online vendors, auction sites, and other corporate users, for which it charges a fee. Private users need to register and set up a profile which, for example, includes a reference to a bank account or to a

credit card; PayPal will use that reference to collect money to be paid by the account or card owner to someone else. When a customer reaches the checkout phase during an e-commerce session, the shopping site he or she interacts with might branch to PayPal for payment processing. If the customer agrees to pay through PayPal, PayPal will verify the payment through a sequence of encrypted messages; if approved, the seller will receive a message stating that the payment has been verified, so that the goods can finally be shipped to the customer. Services such as PayPal have invented the notion of a *micropayment*, i.e., a tiny payment (of often just a few cents) which is feasible as a reimbursement only if occurring sufficiently often.

Other forms of third-party services have become common on the Web, for example companies issuing *digital certificates*. These are files containing the information such as the name to whom it is given, the location (URL) of the respective organization, the public key of that organization, a validity period, and an issuer, and they are issued by certificate authorities such as Verisign or RSA.

In conclusion, it is fair to say that, since the inception of the Web in 1993, a considerable amount of the trading and retail business has moved to electronic platforms and is nowadays run over the Web. This has created completely new industries, and it has led to a number of new business models and side effects that do not exist, at least not at this scale, in the physical world. We will see another impact the evolution of the Web has had next.

2.5 P2P Networks and the No-Cost Mentality

Initially, when the Web started and HTML became available as a markup language for Web pages, people composed their HTML code in a text editor, a way that still works today. A few years later, tools became available for designing Web pages and for setting up Web sites more and more easily. Some of these simply allowed to design HTML documents and to include links, graphics, maybe even audio and video in a WYSIWYG fashion, others even allowed for an easy management of entire Web sites comprised of multiple pages. The modern result of this development are *content management systems* (CMS), which are underneath most major Web sites today, in particular those maintained at an enterprise level.

What is more important to us is the fact that over time more and more people started setting up sites using these tools, and the obvious consequence was that the information available on the Web grew exponentially. Once a site had been created, the next important issue was to get it found, for which the emerging breed of search engines provided registration mechanisms, sometimes for free, increasingly often against a fee. This also led to the development of tricks that, for example, faked high popularity of a site just in order to get a good ranking with search engines. Besides text and textual documents, people soon started to place other types of documents on the Web, in particular media such as image, audio, and video files. Now every user may have experienced how easy it is to save (actually copy) an image found in an HTML document: just right-click on the image and select the “save image as” option! Similarly, audio and video files can easily be downloaded and copied to a local computer, as long as access to these files is granted.

Beyond the Client/Server Model: P2P

It soon turned out, however, that the traditional client/server model behind the Web, which we have described earlier, was not optimal for some interactions, including the downloading of large files. Imagine a video file that contains a 90-minute movie; with a reasonable compression rate this will easily amount to roughly 800 MB in file size. Even over an Internet connection that can guarantee a constant download rate of 80 KB/sec, this will take almost 3 hours! Thus, if a video server would have to serve a request for such a file from a client, the server would be kept busy with just this request for quite a while. Moreover, a *constant* rate can hardly be guaranteed even these days, a fact that has so far prevented video-on-demand from becoming a big business. Clearly, this situation will change as technology advances, yet for the time being an alternative is needed.

Notice that the video download problem is not just a matter of bandwidth; it is also a matter of a single server being occupied with a large request for quite some time. The workaround people have developed for tasks like this are *peer-to-peer* (P2P) networks, which give up the idea of a central server that has to take care of all incoming requests we have outlined earlier. Instead, a P2P network primarily relies on the computing power and bandwidth of its participants and is typically used for connecting nodes via mostly ad-hoc connections [25]. A P2P network also does not know the notion of a client; any participant in the network can function as either a client or a server to the other nodes of the network, as needed by the task at hand. In fact, a P2P system comes with complete and decentralized self-management and resource usage, and it enables two or more peers to collaborate spontaneously in a network of equals (peers) by using appropriate information and communication systems.

As mentioned, one of the many uses for P2P networks is the sharing of large files, which is done on a large scale today on the Internet. The different P2P systems in use are based on distinct file-sharing architectures with different principles, advantages, disadvantages and naming conventions.

Gnutella

Gnutella is a file sharing network that has been set up since 2000 and is among the most popular P2P networks today. Gnutella nowadays has several million users, many of whom might be on at any given time. For using Gnutella, or for being a participant in a Gnutella network, one has to install a Gnutella client on a local machine. From there, requests for files can be issued as illustrated in

Figure 12. Here, a participant has issued a request for a file, which is broadcasted to all other participants in the network. The one holding the file will announce “hit” on the path through the intermediate nodes through which the hit node was found; the subsequent download will then go directly to the requester.

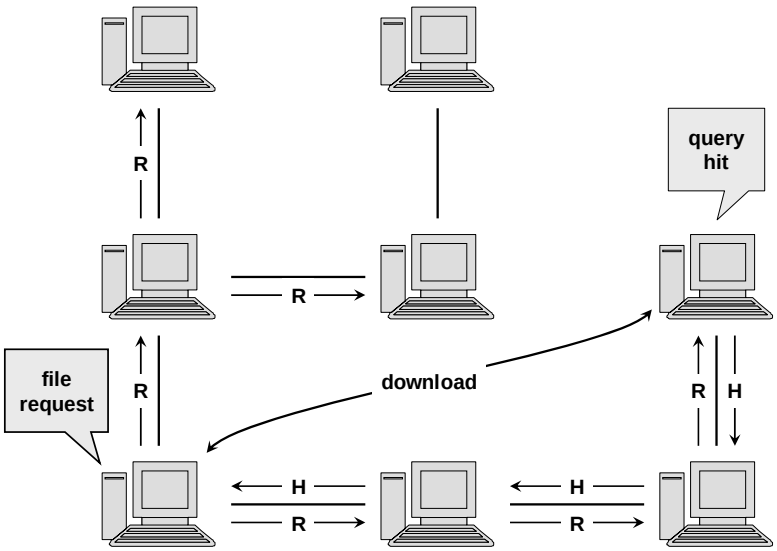

Figure 12: File Request in a Gnutella Network.

Notice the absence of a central server, which is typical for a P2P network. Control is entirely decentralized. A search query or request is forwarded in a Gnutella network until the TTL (time to live) of the request packet has become zero. The number of queries that are issued during a request may rise exponentially with the number of peers that receive such a query, making the network rather inefficient but resistant to failures.

There are other popular P2P networks that have different approaches to certain parts of this basic model, mostly to circumvent the exponential rise of requests that Gnutella faces.

Open Access

One of the consequences of developments like the P2P networks, protocols, and tools just described has always been that information and files started to become available on the Web which previously had been pretty costly. We are not delving into legal issues related to services like Napster and Kazaa here or into general issues related to copyright or intellectual rights and their protection. However, fact is that many users around the globe have started using the Internet and the Web as a free source for almost everything. For example, once the mp3 format had been invented as a digital audio encoding and lossy compression format by Fraunhofer's *Institut für Integrierte Schaltungen* in Erlangen, Germany, music got transformed into mp3 format and then could be copied freely between computers and other devices. On the illegal side, people started ripping music CDs and exchanging their content over the Internet; others took videos of recently released movies with a camcorder in a cinema, compressed them into a suitable video format, and put them on a file-sharing network for general copying.

A net result of this is a perception of the Internet and the Web nowadays, shared by many of their users (in particular the younger ones), that all content is, or should be, for free. As we have tried to point out, the issues are far from being resolved even today, but fact is that sharing stuff for free, be it legal or not, has become extremely popular. Open sharing, on the other hand, also has positive effects, as can be seen from the open source community of software developers

(see Friedman's flatteners above). *Open access* at no charge and without license restrictions in particular to scientific results and literature has become a radical demand by many people these days and has even emerged into an international movement. The idea is already supported through a growing number of electronic journals which publish scientific papers exclusively in electronic form and often for free; it is also supported by numerous university sites world-wide, where scientists put their papers, talks, and other material up for free download. On the other hand, it is clear that a publishing company is unable to provide all its copyrighted material, in particular books, for free. The discussion on this topic is in full swing these days, without a conclusion that would be acceptable to all participants in sight.

Free Services after Registration

Many sites on the Web have created a new form of usage for the no-cost mentality that has emerged from the spread of P2P networks, and have even found a new way of doing business with it. Indeed, a number of sites nowadays will say to their visitors something like "register, and from then on the services available from my site are for free." Prominent examples include MySpace.com, Yahoo! and Google. Once you have registered for an account at any of these or many other sites, you may be allowed to use storage space, upload files, invite other participants to access your files, use their free email service etc. Moreover, it is often even possible to *personalize* your site, i.e., to customize pages according to your preferences, desires, or needs. What you may have to accept as a kind of compensation is that advertisements will be placed on your opening page (or even beyond that) or be sent to your email account from time to time. As we have mentioned earlier, advertising on the Web has become one of the most prominent Internet business models, and the idea of "free" sites just described provides yet another advertising channel. Clearly, the more people register at a site, i.e., reveal some of their personal data and maybe even a "user profile" of preferences and hobbies, the more data the site owner will have available and the more he can do with it. Experience also shows that people do not re-register too often; in other words, if you have gone through some registration process once, people often hesitate to do again over and over. Thus, there is some form of "customer retention" right away, and then is it often just a small step to starting to offer these customers a little extra service for which they then, however, have to pay.

With these remarks we conclude the applications stream. The bottom-line we can take away from the preceding discussion is that the Internet and in particular the Web have penetrated our lives in a number of (primarily positive) ways (even though some people have gotten addicted to it, and even shrinks world-wide have found numerous new customers due to the rise of the Web). Search is a primary activity on the Web today, e-commerce and electronic banking have become big business, and new models for doing business such as C2C auctions or for presenting advertisements have emerged. The reader is referred to Pogue and Biersdorfer [22] for further details on many of the topics we have touched upon. All of this has been made possible by a number of technological advances, both on the hardware and on the software side, as will be discussed next.

3 Technology Advances in a Nutshell

We have already mentioned some of the (hardware or software) technology that has emerged in the context of the Web, that has helped to make the Web popular, or that serves as a foundation in general. We now consider the technology stream of a development that has finally led to what is now being called “Web 2.0” in a little more detail and touch upon the most relevant technological advances and communication infrastructure in this section. However, this is not intended as an in-depth treatment of hardware, networking, or software technology (which is the order in which we will discuss it), so we again refer the reader for details to the relevant literature.

Hardware

In hardware, there is essentially one singular development that governs it all: the fact that hardware is becoming smaller and smaller and will ultimately disappear from visibility. Consider, for example, the personal computer (PC). While already more than 10 years old when the Web was launched, it has shrunk (and become cheaper) on a regular basis ever since, with laptops being more popular (and higher in sales figures) than desktops nowadays. Moreover, with processors embedded into other systems such as cars, cell phones, watches etc., we can now carry a computing power in our pockets that was unthinkable only a few years back (and that typically outperforms the computing power that was needed to fly man to the Moon in the late 1960s). Just think of an intelligent cell phone or a modern PDA that is powered by a microprocessor, has some 64 MB or at least 32 MB of main memory, maybe even a hard drive, that can run a slightly simplified version of a common operating system, and that can have a host of application installed (and can have many of them running simultaneously). Thus, in many applications we do not even see the computer anymore, and this trend, which has been envisioned, for example, by Donald Norman in [17], will continue in ways we cannot even fully predict today.

Another important aspect of hardware development has always been that prices keep dropping, in spite of expectations that this cannot go on forever. As has been noted recently, Moore’s Law is still valid after 40 years, and expected to remain valid for another 10 to 15 years. In this “law” Gordon Moore, one of the founders of Intel, predicted in 1965 in an article in the journal “Electronics” that the number transistors on a chip would double every 12 to 18 months; he later corrected the time span to 24 months, but that does not change anything significant. It turns out that microprocessor packaging has vastly been able to keep up with this law, and no reason is in sight why this should change soon. Raymond Kurzweil, one of the primary visionaries of artificial intelligence and father of famous music synthesizers, and others consider Moore’s Law a special case of a more general law that applies to the technological evolution in general: If the potential of a specific technology is exhausted, it is replaced by a new one. Kurzweil does not use the “transistors-per-chip” measure, but prefers “computing power per $1,000 machine.” Indeed, considering the evolution of computers from mechanical devices via tubes and transistors to present-day microprocessors, it exhibits a double-exponential growth of efficiency. The computing power per $1,000 (mechanical) computer has doubled between 1910 and 1950 every three years, between 1950 and 1966 roughly every two years, and nowadays doubles almost annually.

As a result, hardware has become a commodity, cheap and ubiquitous. For being able to use hardware, be it processors or storage, it is not even necessary anymore to purchase it, since computing power and storage capacity can nowadays be rented on-demand. With many Internet providers (as well as other companies, for example Amazon.com with its S3 service) private or commercial customers can choose the type of machine they need (with characteristics such as number of processors, clock frequency, or main memory), the desired amount of storage, and the rental period and then get charged, say, on a monthly basis. This has even become an attractive alternative to purchasing for the reason that, since hardware ages so fast, there is no more need to get rid of items no longer used.

In the academic world, the idea of *grid computing* [8] has emerged as an alternative, where applications can request computational power in an on-demand fashion within an always available "grid" of machines (and often use them for free), just like electric power is obtained from an outlet, but is actually drawn from an electrical grid that does not need any user interaction or adaptation even at peak times. Grid computing has a number of applications, for example in particle physics, meteorology, medical data processing or in satellite image processing. It has become widely known through the SETI@home project of the University of California, Berkeley, which uses computer idle time for a search for extraterrestrial intelligence (http://setiathome.berkeley.edu/).

IP Networking

Interestingly, a similar trend of evolving from an expensive and rare technicality into a cheap and ubiquitous commodity can be observed regarding computer networks, at least from an end-user perspective. This is especially true for networks that are based on the TCP/IP protocol stack and which are hence considered to be part of the Internet; for details of TCP (Transmission Control Protocol) and IP (Internet Protocol), we refer the reader to Tanenbaum [25]. In particular the arrival of wireless network technology has made it possible to get connected to the Internet without cables, and many modern devices are able to establish a connection to the nearest hot spot just by themselves. At the same time, cable-based networks, with fiber optics having replaced copper wires to a large extent (again an enabler of the flattening of the world mentioned earlier), start moving into private homes, as the technology continues to decrease in price, and convergence of technologies is on the horizon. For instance, some providers nowadays let users get an Internet connection over a powerline; others integrate Internet and telephone communications. The latter has become known under acronym *Voice over IP* (VoIP) and has been made popular especially by *Skype*. Skype users make telephone and video calls through their computer using Skype client software and an Internet connection. Users may also communicate with landline and mobile telephones, although this requires setting up an account in which the caller has deposited money. Skype operates on a P2P model rather than a client-server model. The Skype user directory is entirely decentralized and distributed among the nodes in the network, so that the network can easily scale to large sizes.

As Michael Roberts notes in [24], the historical development of the Internet can be seen in four stages: Stage 1 refers to the period from 1980 to 1991, where the net had a research and academic focus; Stage 2 is 1992 to 1997, which saw early public uses; Stage 3 followed from 1998 to 2005, when the Internet achieved both domestic and international critical mass; we are currently experiencing the fourth stage. Especially Stage 3 was impacted considerably by the

aforementioned dot-com bubble and its gigantic speculation in Internet stocks, which laid the foundation for broadband Internet applications and the integration of data, voice, and video services on the single technological basis that we are used to today. As remarked also by Friedman [10], one of the consequences of the burst of the dot-com bubble was an oversupply of fiber-optic cable capacity especially in the US, of which many newly created service providers were able to take advantage.

The mid-1990s also saw a growing need for administration of Internet issues, one result of which was the creation of ICANN, the *Internet Corporation for Assigned Names and Numbers*. ICANN is a private non-profit organization based in Marina del Rey, California, whose basic task is the technical administration of the Internet, in particular the assignment of domain names and IP addresses as well as the introduction of new top-level domains. To this end, it is worth mentioning that naming on the Internet follows a hierarchical pattern as defined in the *Domain Name System* (DNS), which translates domain or computer hostnames into IP addresses, thereby providing a worldwide keyword-based redirection service. It also lists mail exchange servers accepting e-mail for each domain, and it makes it possible for people to assign authoritative names without needing to communicate with a central registrar each time. To give the reader an idea of how many names are around these days, in the fall of 2006 there were more than 53 million .com domains registered, more than 10 million .de domains, or almost 5 million .org domains. The mid-1990s moreover saw the formation of the *World Wide Web Consortium* (W3C), the main international organization that deals with the development of standards related to Web technology. It was founded by Web inventor Tim Berners-Lee.

The Broadband Era

With broadband and wireless technology as a commodity (with the only major exception being the developing countries), we will soon see a host of new applications and services arise and delivered over the Internet and the Web, with digital radio and television only being precursors of what is there to come. Broadband communication in particular allows for an easy transfer of large files, so that, for example, it becomes possible to watch movies over the Internet on a mobile device, since at some point it will be possible to guarantee a constant transfer rate over a certain period of time. As Figure 13 indicates, the broadband penetration of homes in the US has gone up considerably since the year 2000. A typical effect after getting broadband at home is that people spend more time on the Internet. Moreover, with flat rate for Internet access nowadays widely available, many users do not explicitly switch their connection on and off, but are essentially “always on.”

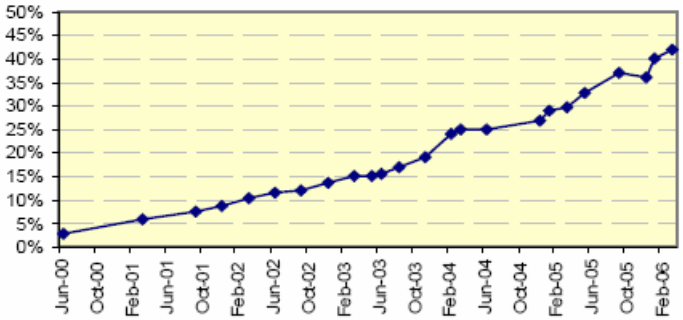

Figure 13: Home Broadband Penetration in the US.
Source: *http://www.pewinternet.org/pdfs/PIP_Broadband_trends2006.pdf*

HTML and XML

On the software side, we first have to mention that in parallel to hardware becoming a commodity, software development has dropped in prices, since what has been an expensive service some time ago is now cheap, and the "LAMP" (a common acronym for *Linux, Apache, MySQL, PHP*) manifestation of the open-source world has brought along a considerable number of tools through which software development is nowadays supported.

We have already mentioned the arrival, together with the Web, of a new language, HTML, the predominant markup language for the creation of Web pages. HTML provides means to structure text-based information in a document by denoting headings, tables, paragraphs, or lists, and to supplement that text with forms, images, links, and interaction. The language was originally developed by Tim Berners-Lee in the context of his creation of the Web, and it became popular through the fact that it is easy to use. An HTML document can quickly be set up using just a few structuring elements again called *tags* (see for example [16]). Tags have to follow some simple syntactical rules and are often used to describe both content and presentation of a document. The *separation* of presentation and content became an issue when Web pages started to be rendered on more and more devices, including computer terminals, laptop screens, and cell phone displays, since each device has its own capabilities, requirements, and restrictions. In HTML, presentation can be specified within a document or separately within a *cascading style sheet* or CSS files.

HTML tags are all predefined, and although there are ways to include additional tags (for example, through the embedding of scripting language code), tags can generally not be user-defined. This is different in XML [11], the *Extensible Markup Language*, a W3C recommendation for a general-purpose markup language that supports a wide variety of applications. Markup languages based on the XML standard are easy to design, and this has been done for such diverse fields as astronomy, biochemistry, music, or mathematics and for such distinct applications like voice or news in recent years. XML-based languages are also reasonably human-readable, since the tags used can be chosen in such a way that they relate to the meaning of the particular portion of the document that they enclose. XML is a simplified subset of the *Standard Generalized Markup Language* (SGML) and is widely used in information integration and sharing applications, in particular as they arise on the Internet. Any XML-based language should have an associated syntax specification, which can take the form of a *document type definition* (DTD, a formalism essentially based on regular context-free

languages, or of an *XML Schema Definition* (XSD), which specifies a schema roughly in the style and detail of structure and type declarations found in programming languages or database schema languages.

The development of a new language based on XML typically goes through several steps, including the following:

1. Search for existing DTDs (or schemas);
2. Design of a namespace and a DTD (or a schema);
3. Discussion and fine tuning within the interested community;
4. Publication of the language;
5. Community-wide use of the language.

The concept of a *namespace* has turned out to be instrumental for XML. Namespaces provide a simple method for providing uniquely named elements and attributes in XML documents, namely by associating them with a distinct name collection that is identified by a URI. An XML document may contain elements from different namespaces, as long as they are all properly referenced in the beginning of the document.

There are pros and cons regarding the question of whether to use a DTD or an XML Schema, but such a discussion goes beyond the scope of this report. XML has had an impact on HTML in that it has brought along XHTML [16], a version of HTML that follows the same strict syntax rules as XML. More importantly, XML has become a universal enabler for a number of applications on the Web. For example, e-commerce sites use XML-based language intensively for document exchange or integration. Examples include RosettaNet (www.rosettanet.org), and ebXML (Electronic Business using eXtensible Markup Language, see www.ebxml.org), platforms which provide standardized XML documents for e-commerce items such as orders, invoices, etc. RosettaNet is an open, non-profit consortium of industrial companies that develops universal standards for the global supply chain. ebXML was started in 1999 as an initiative of OASIS (Organization for the Advancement of Structured Information Standards) and the United Nations/ECE agency CEFACT and is a collection of specifications that enables arbitrary enterprises to conduct business over the Internet; it provides standard methods for exchanging business messages, for conducting trade relationships, for communicating data in common terms, or for defining and registering business processes.

Beyond Plain HTML: Scripting

As mentioned above, an HTML document is allowed to have scripting code embedded. This arose out of the necessity to make Web pages dynamic as well as interactive. Indeed, often when users are asked for input, that input needs to be checked for correctness and completeness, or it needs to be sent off to a server for verification and most likely storage in a database. Moreover, the response a Web server creates upon the arrival of user input may have to be generated dynamically, e.g., to acknowledge or reject the input, in which case HTML needs to be created on the fly.

To this end, an important distinction refers to the question of whether scripting occurs at the client side or at the server side. Consider the typical Web client/server scenario shown in

Figure 14 where the client basically runs a browser, while the server has access to a number of stored documents and data sources.

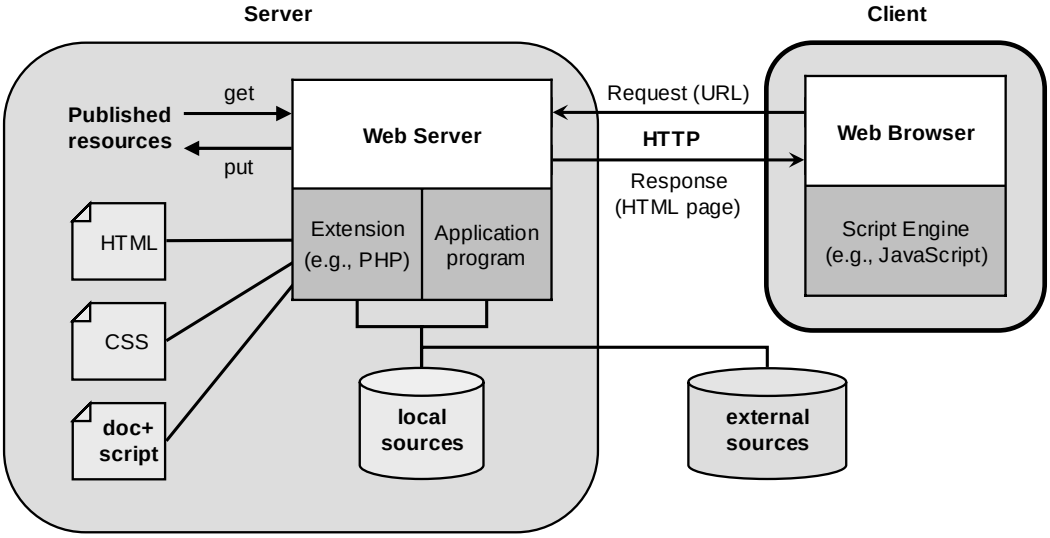

Figure 14: Cient-side vs. Server-side scripting.

Client-side scripting, very often seen in the form of *JavaScript*, makes use of the fact that a browser can not only render HTML pages, but also execute programs. These programs, which have to be written in a script language, will be interpreted just like HTML code in general. Thus, some of the tasks arising in a Web site can be off-loaded onto the client. On the other hand, certain things cannot be done at the client side, in particular when access to a database on the Web is needed. With server-side scripting using, for example, the *PHP* language, user requests are fulfilled by running a script directly on the Web server to generate dynamic HTML pages; it can be used to provide interactive Web sites that interface to databases or other data stores as well as local or external sources, with the primary advantage being the ability to customize the response based on a user's requirements, access rights, or query results returned by a database.

All of the above, both client-side and server-side scripting, is based on the classical client/server paradigm we have mentioned earlier and on the fact that any such interaction so far has been assumed to be *synchronous*. In order to enhance Web programming even further, a recent idea has been to not only allow HTML creation or modification on the fly (“dynamically”), but to be able to provide direct feedback to the user via on-the-fly HTML generation on the client. This, combined with *asynchronous* processing of data which allows sending data directly to the server for processing and receiving responses from the server *without* the need to reload an entire page, has led to a further separation of user interface logic from business logic now know under the acronym AJAX (*Asynchronous JavaScript and XML*). In a nutshell, AJAX brings together the various (software) technologies shown in Figure 15.

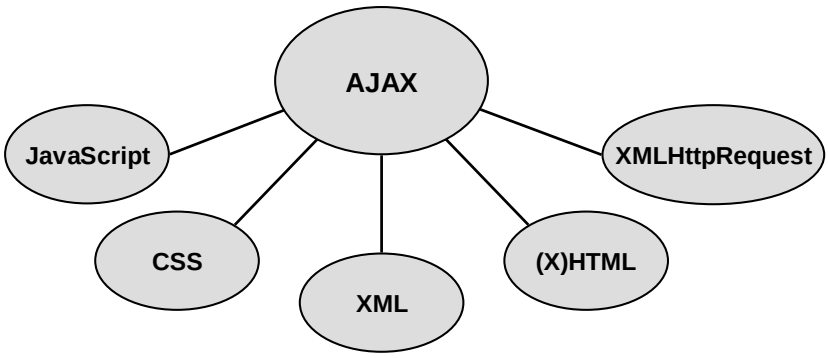

Figure 15: AJAX and its constituents.

AJAX is a Web development technique that allows developers to build “rich” Web applications that are similar in functionality to classical desktop applications, yet they run in a Web browser. Its main functionality stems from an exploitation of XMLHttpRequest, a JavaScript class (with specific properties and methods) supported by most browsers which allows HTTP requests to be sent from inside JavaScript code.

More XML Applications

Out of the numerous applications XML has seen to date, we here mention just two more: Web services and RSS. Web services extend the client/server paradigm by the notion of an explicit repository or registry as indicated in Figure 16, thereby solving the problem of locating a service in a way that is appropriate for the Web.

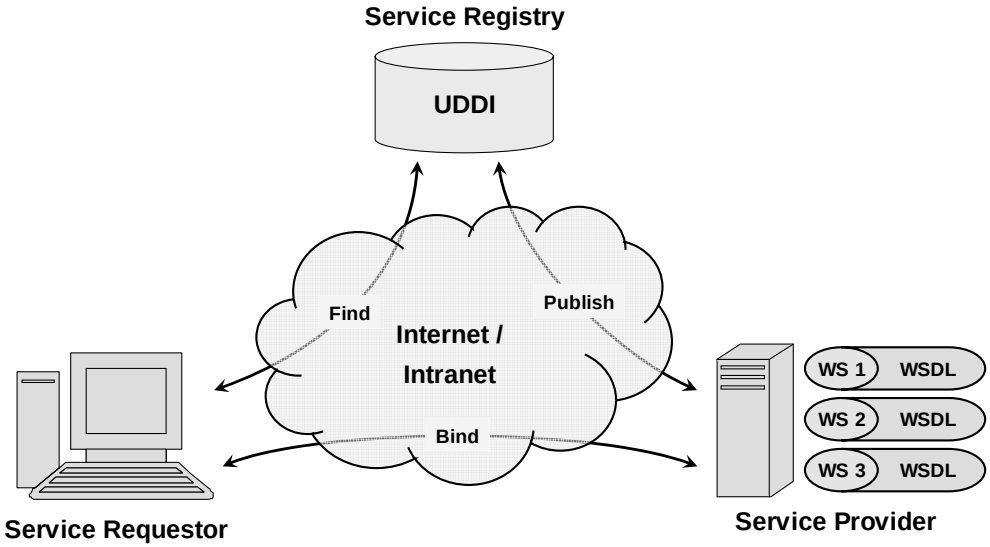

Figure 16: The Web service paradigm.

A service requestor (client) looking for a service sends a corresponding query to a service registry or repository. If the desired service is found, the client can contact the service provider

and use the service. The situation and proceeding is similar to looking for a service in real life by consulting the local Yellow Pages for alternative offering and contact data. The provider has previously published his service(s) in the registry. Hence, Web services hide all details concerning their implementation and the platforms they are based on; they essentially come with a unique URI (the *Uniform Resource Identifier* is a unique address) that points to their provider. Since Web services are generally assumed to be interoperable, they can be combined with other services to build new applications with more comprehensive functionality than any single service involved.

To achieve these goals, Web services are commonly based on standards, in particular on the XML-based specifications SOAP (*Simple Object Access Protocol*), UDDI (*Universal Description, Discovery and Integration*), and WSDL (*Web Services Description Language*), all of which come with their distinct namespace and are described in detail, for example, in [1]. These standards have typically been made a recommendation by the W3C (or are in the process of becoming a recommendation, which sometime takes a bit of time). The benefits of a Web services architecture is well recognized in the business-to-business (B2B) area already, where companies are starting to use it for enterprise application integration, B2B integration, application construction and a flexible approach to outsourcing. What is on the horizon today is a massive exploitation of the service paradigm to software provisioning even for customer applications. As we will see, modern Web applications tend to offer more and more services without the need of software installation, yet not every single one of them follows the paradigm shown in Figure 16 literally.

The second major XML application we mention here marks a departure from the traditional way of looking for information on the Web, namely by searching and then essentially “pulling” the information found off the Web. As people have done for decades with newspapers and magazines, a more convenient way to obtain information on a regular basis is to *subscribe* to it, an idea underlying *push services* on the Web. Here the object in question is the *Web feed*, a data format used for pushing frequently updated content to interested users. Content distributors “syndicate” Web feeds, thereby allowing users to subscribe to it. Thus, in a typical usage scenario for such feeds, a content provider publishes a feed link on his site; users can subscribe to this using a feed or news reader running on their own machines. When instructed or periodically, the reader asks all the servers in its feed list if they have new content; if so, it either makes a note of the new content or immediately starts downloading it (though strictly speaking this again results in a “pull”, we call it *push* here as the user does not have to activate this each time).

The primary format in which these feeds are specified is XML-based RSS, which may refer to *Really Simple Syndication* (RSS 2.0), *Rich Site Summary* (RSS 0.91, RSS 1.0), or *RDF Site Summary.* Using RSS or the Atom syndication format, users or clients can subscribe to a host of information from the Web, so that obtaining information (e.g., stock tickers, special offers, news, and nowadays often *podcasts*) can even be automated in various ways. RSS and Atom have provided in some form what is nowadays called a *mashup* for several years already.

This concludes our brief tour through some of the most relevant technological advances in connection with the Internet and the Web over recent years. There has been enormous progress in networking technology and in hardware, and there have also been a number of good ideas in the software area, which together have helped to make the Web an extremely

popular and widely used medium. What we look at in the next subsection is how user perception of that medium as well as user interaction with it have changed drastically within a relatively short period of time.

4 User Perception and Participation: Acceptances of the Web

The third preparatory stream for Web 2.0 we will discuss here is intended to show how different generations of people have changed their perception of and participation in the Web, and have accepted the Web as a communication medium, as a discussion forum, as a business platform, as a storage device even for their diaries, or as a constantly growing and expanding encyclopedia. To this end, imagine an average 18-year old student back in 1993. This person probably did not even notice that the Web came along. Clearly, there were others who did, but for the average person it took a few years to recognize what the Web could do for her or him. The same person, now in his or her early 30s, will meanwhile have gotten used to electronic mail, both in his or her business and private life, bank account and portfolio management, search for arbitrary information, exchange photos with friends, and probably a few other things that can easily be done over the Web today. What we try to summarize in this section is how a person who is 18 years of age at the end of 2006 will perceive the Web, which typically is radically different.

4.1 Blogs and Wikis

We have mentioned occasionally throughout this report that users have started to use the Web as a medium in which they can easily and freely express themselves, and by doing so online they can even reach a high number of other people most of which they will not even know. We will now take a closer look at two forms that have become pretty popular, and that are used today for completely different purposes: Blogs are typically expressions of personal or professional opinion or experience which other people can at most comment, while wikis are pages or systems of pages describing content that other people can directly edit and hence extend, update, modify, or even delete. Both communication forms have become very popular in recent years and contribute significantly to the novel nature of the Web today.

Blogs and New Forms of Publishing

An effect the Web has seen in recent years and that we have described above is that people have started to write comments on products or sellers, on trips or special offers, and more generally nowadays on almost any topic; ultimately people have started to even write about themselves, or have started to comment on vastly any issue even without a particular cause (such as a prior shopping experience). This had led to the creation of “blogs” and to a new form of activity called “blogging.” In essence, a *blog* is an online diary or a journal that a person is keeping and updating on an ad-hoc or a regular basis. The word itself is a shortened version of *Web log* and is meant to resemble the logs kept by the captain of a ship as a written record of daily activities and documentation describing a journey of the ship.

A blog on the Web is typically a sequence of short texts that is organized like what computer scientists call a “stack,” i.e., entries appear in reverse order of publication so that the most recent entry is always shown first. In its most basic form, a blog consists of text only. Without any additional features and in particular if separated from subscriptions and from RSS, a blog is hence no more than a kind of diary that may be kept by anybody, e.g., private persons, people

in prominent positions, politicians, movie stars, musicians, companies, or even company CEOs. However, most blogs go way beyond a simple functionality today.

As a first example of a blog, let us consider Slashdot (www.slashdot.org). It was started in 1997 by Rob Malda for publishing "news for nerds, stuff that matters." He still maintains the blog today and has created one of the most lively sites for Linux kernel news, cartoons, open-source projects, Internet law, and many other issues, categorized into areas such as *Books, Developers, Games, Hardware, Interviews, IT, Linux, Politics,* or *Science*. Each entry is attached to a discussion forum where comments can even be placed anonymously. A second example of a blog, which also exhibits an interesting idea of how to use this medium, can be found at http://www.historymatters.org.uk and represents an attempt to motivate thousands of people in the U.K. to contribute to the largest blog in history, by writing about what they have been doing on October 17, 2006.

As a prominent example from a company, Figure 17 shows an excerpt from the blog kept by General Motors Vice Chairman Bob Lutz, in which he and other GM people report on new developments or events in any of the GM companies (e.g., the presentation of the new Buick Enclave at the 2006 LA Auto Show or the interior design of the new Chevrolet Malibu), respond to consumer enquiries, or present a link to a video that can be found on GMtv. The blog is not a simple one, since it comes with a number of features: Users can subscribe to it, i.e., be notified when new entries get posted, existing blog entries are placed in different categories, users can access an archive that goes at least two years back, and there are many links to other sites inside and outside GM. The blog is hence a good example of how an enterprise can develop new strategies for its external (but also its internal) communication.

There is even a list of other blogs on Bob Lutz's blog, such as the Dilbert blog (http://dilbertblog.typepad.com/). The latter also exhibits typical features found in blog sites: While blogging services are often for free, i.e., users can create and maintain a blog of their own with any charges, they typically have to accept advertising around their entries. The Dilbert blog is maintained by Typepad, which hosts free blogs only for a trial period, but which explains a number of aspects why people actually do blog, namely to broadcast personal news to the world, to share a passion or a hobby, to find a new job, or to write about their current one. Moreover, companies like GM have discovered that blogs can be a medium through which they can link with customers and maybe even get new ones interested in buying their products.

Blogs are one of several modern forms of writing to the Web. Providers where a blog can be set up (typically within a few minutes) include Blogger (www.blogger.com), Blogging (www.blogging.com), Yahoo! 360° (360.yahoo.com), or LiveJournal (www.livejournal.com). UserLand was among the first to produce professional blogging software called Radio (radio.userland.com). If a blog is set up with a provider, it will be often the case that the blog is immediately created in a structured format such as RSS so that readers of the blog will be informed about new entries.

The activity of blogging, which is often enhanced, for example, with images, audio or even video podcasts, can be seen as the successor to bulletin boards and forums, which have existed on the Internet roughly since the mid-90s. Their numbers have increased enormously in recent years, which is why the advertising industry is taking a close look at them. And in a similar way as has commenting on products on commerce sites or evaluating sellers on auction sites,

blogging is starting to contribute to consumer behavior, since an individual user can now express his or her opinion without someone else executing control over it. Clearly, the party hosting a blog has an ethical responsibility and can block a blog or even take it off-line, yet people can basically post their opinions freely. Studies are showing that trust in private opinion is generally high. Blogs may also be moderated which applies, for example, to the aforementioned Slashdot; this site is also a good example for the influence a blog may have on the public or politics or both, demonstrated for example by the story of the Russian programmer Dmitri Sklyarov (see http://slashdot.org/yro/01/07/20/1332227.shtml).

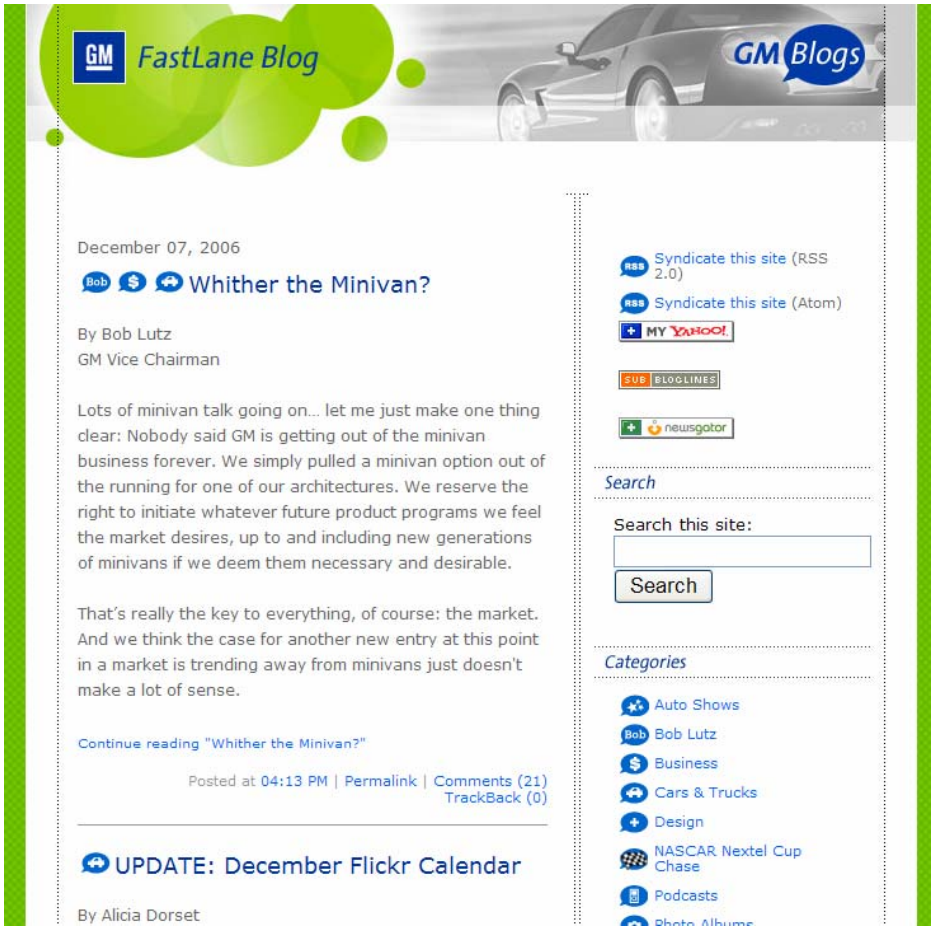

Figure 17: Bob Lutz's blog on auto news from Detroit.
Source: *http://fastlane.gmblogs.com*

Blogs have also been discovered by search engines and are nowadays visited by crawlers on a regular basis. Since blogs can contain links to other blogs and other sites on Web and links can be seen as a way for bloggers to refer to and collaborate with each other, and since link analysis mechanisms such as Google's PageRank give higher preference to sites with more incoming links, bloggers can even influence the ranking of sites at search engines. Besides search engines like the ones we have mentioned, more specialized ones have meanwhile been developed for searching through or for blogs, including Daypop (http://www.daypop.com/) and Technorati (www.technorati.com).

Obviously, blogging has also opened the door for new forms of misuse. For example, blogs or blog entries can be “requested” in the sense that people are asked to write nice things about a product or an employer into a blog, and they might even get paid for this. Conversely, a blog could be used to mob a person or a product. An easy way to avoid some of these misuses is to require that blog writers have to identify themselves. Nevertheless, blogs have become a highly popular medium for people to express themselves; readers interested in learning more about the topic are referred, for example, to [29].

Wikis and Wikipedia

Another prominent form of user participation on the Web today is represented by wikis. A *wiki* is a Web page or a collection of pages that allows its users to add, remove, and generally edit and modify some of the available content, sometimes even without the need for prior registration if the wiki is a public one. Thus, a wiki is an editable page or page collection that does not even require its users to know how to write a document in HTML. The term “wiki” is derived from the Hawaiian word “wikiwiki” which means “fast.” Thus, the name suggests having a fast medium for collaborative publication of content on the Web. Sometimes a distinction is made between a single “wiki page” and “the wiki” as an entire site of pages that are connected through many links and which is in effect a simple, easy-to-use, and user-maintained database for creating content.

The history of wikis started in March 1995, when Ward Cunningham, a software designer from Portland, Oregon, was working on software design patterns and wanted to create a database of patterns in which other designers could contribute by refining existing patterns or by adding new ones. He extended his already existing “Portland Pattern Repository” by a database for patterns which he called *WikiWikiWeb*. The goal was to fill the database with content quickly, and in order to achieve this, he implemented a simple idea which can be seen in Figure 18: Each page has at its bottom a link entitled “EditText” which can be used to edit the text in the core of the page directly in the browser. For example, pressing the edit link in the page shown in Figure 18 will lead to the page shown in Figure 19.

As can be seen, users can now write within the bordered area shown, and they can save their edits after entering the code number seen above this area. The important point is that no HTML is needed to edit the page; instead, the new or modified content will be converted to HTML by appropriate wiki software. Some rules apply in case a user wants to input headlines, emphasized words etc.

Software Design Patterns

Most of the discussion of DesignPatterns on this Wiki are specific to SoftwareEngineering and are therefore SoftwareDesignPatterns, as contrasted with Alexander's original ArchitecturalDesignPatterns, which inspired the ones in software.

SoftwareDesignPatterns include famous examples such as **SingletonPattern**, **AbstractFactoryPattern** (see **BuilderPattern**), **PrototypePattern**

SoftwareDesignPatternsIndex with cross references to different names for the same pattern.

See CategoryPattern for a moderately exhaustive list.

Somewhat related topics:

- **ExtremeProgramming**
- **UnitTests**

EditText of this page (last edited June 2, 2005)
FindPage by searching (or browse LikePages or take a VisualTour)

Figure 18: Software Design Pattern Wiki in presentation mode.
Source: *http://c2.com/cgi/wiki?SoftwareDesignPatterns*

SoftwareDesignPatterns

Advice to visitors: Spam is not allowed on this site. Unwanted links are removed before indexing is allowed. If you are new here, please consider reading GoodStyle before contributing. If you just want to try out how Wiki works, please edit WikiWikiSandbox instead of existing pages or adding new ones. Thank you.

Type the code word, 567, here [ ] then press [Save] to finish editing. Read MoreAboutCodes.

```
Most of the discussion of DesignPatterns on this Wiki are specific to SoftwareEngineering and are
therefore SoftwareDesignPatterns, as contrasted with Alexander's original ArchitecturalDesignPatterns,
which inspired the ones in software.

SoftwareDesignPatterns include famous examples such
as '''SingletonPattern''', '''AbstractFactoryPattern''' (see '''BuilderPattern'''), '''PrototypePattern'''

SoftwareDesignPatternsIndex with cross references to different names for the same pattern.

See CategoryPattern for a moderately exhaustive list.
----

Somewhat related topics:
*'''ExtremeProgramming'''
*'''UnitTest''''''s'''
```

☐ I can not type tabs. Please ConvertSpacesToTabs for me when I save.

GoodStyle tips for editing.
EditPage using a smaller text area.
EditCopy from previous author.

Figure 19: Software Design Patterns Wiki in editing mode.
Source: *http://c2.com/cgi/wiki?edit=SoftwareDesignPatterns*

Often, there is no review before the modifications a user has made to a wiki page are accepted. Commonly, edits can be made in real-time, and appear online immediately. There are systems that allow or even require a login-in which allows signed edits, through which the author of a modification can be identified. In particular, log-in is often required for private wiki servers only after which a user is able to edit or read the contents. Most wiki systems have the ability to record changes so that an edit can be made undone and the respective page be brought back

into any of its previous states. They can also show most recent changes and support a history, and often there is a “diff” function that helps readers to locate differences to a previous edit or between two revisions of the same page. As with blogs, there is an obvious possibility to abuse a wiki system and input garbage.

There are currently more than 100 systems for creating and maintaining wikis, most of which are open source (see http://c2.com/cgi/wiki?WikiEngines for a pretty up-to-date listing). A prominent example is Wikia (www.wikia.com), Wikia was founded by Angela Beesley and Jimmy Wales, the latter of which also created Wikipedia, under the name “Wikicities”, in October 2004 and relaunched as “Wikia” in March 2006. It also thus does not surprise that there are meanwhile wikis for almost each and every topic, a prominent example being *The Hitchhiker’s Guide to the Galaxy* (http://www.bbc.co.uk/dna/h2g2). This wiki was created by *The Digital Village*, a company owned by author Douglas Adams who also wrote the famous book by the same title; it is nowadays run by the BBC which took it over after Adams’ untimely demise. Adams created it in an attempt to realize his vision of an open encyclopedia authored solely by its users.

The largest wiki today is definitely the multi-lingual Wikipedia project which contains several million articles in about 100 languages and a large number of pictures. Wikipedia started out as an experiment in 2001 to create the largest online encyclopedia ever, but was soon growing faster than most other wikis. It soon got mentioned in blogs and various print media including *The New York Times*. Wikipedia has gone through several cycles of software development, and has always strictly separated content from comments and from pages about Wikipedia itself. Wikipedia has many distinct features that make its use transparent, among them instructions on citations, which anybody who wants to refer to a Wikipedia article elsewhere can easily download. For example, the bibliographic details for entry “wiki” look as shown in Figure 20.

The Wikipedia administration has established pretty strict rules regarding content or users. For example, a page to be deleted must be entered into a “votes for deletion” page, where users can object to its deletion; this reflects the Wikipedia idea of making decisions vastly uniform. By a similar token, a user may only be excluded from Wikipedia right away in a case of vandalism, which has turned out to be rare. In general, discussions that take place about the content of articles are most of the time highly civilized, and Wikipedia has become prototypical for proving that the ease of interaction and operation make wikis an effective tool for collaborative authoring.

Bibliographic details for "Wiki"

Page name: Wiki

Author: Wikipedia contributors

Publisher: Wikipedia, The Free Encyclopedia.

Date of last revision: 11 December 2006 01:06 UTC

Date retrieved: 11 December 2006 01:45 UTC

Permanent link: http://en.wikipedia.org/w/index.php?title=Wiki&oldid=93486004

Page Version ID: 93486004

Please remember to check your manual of style, standards guide or instructor's guidelines for the exact syntax to suit your needs. For more detailed advice, see **Citing Wikipedia**.

Figure 20: Bibliographic details for Wikipedia entry „Wiki“.

How effective especially Wikipedia indeed is came out in a study the famous British science journal *Nature* reported upon in December 2005. In an article entitled *Internet encyclopaedias go head to head, Nature* wrote that “Jimmy Wales' Wikipedia comes close to Britannica in terms of the accuracy of its science entries, a Nature investigation finds” (http://www.nature.com/news/2005/051212/full/438900a.html). For this study, Nature had chosen articles from both Wikipedia and the Encyclopaedia Britannica in a wide range of topics and had sent them to experts for peer review (i.e., without indicating the source of an article). The experts compared the articles one by one from each site on a given topic. 42 of the returned reviews turned out to be usable, and *Nature* found just eight serious errors in the articles, of which four came from either site. However, the reviewers discovered a series of factual errors, omissions, or misleading statements; in total Wikipedia had 162 of them, while the Encyclopaedia Britannica had 123. This averages to 2.92 mistakes per article for the latter and 3.86 for Wikipedia.

The accuracy of Wikipedia may in part be due to the fact that its community has developed several methods for evaluating the quality of an article, including stylistic recommendations, tips for writing good articles, a “cleanup” listing articles that need improvement, or an arbitration committee for complex user conflicts, where these methods can vary from one country to another; some articles even show in their header that they “need improvement.”. The bottom-line is that Wikipedia is presently one of the best examples for an online community that, in spite of permanent warnings such as [7], works extremely well, that has many beneficiaries all over the world, that is in wide use both online and off-line (through DVD distributions that are made available from time to time), and that even enjoys a high degree of trust. Wikipedia also is an excellent example of a platform that is “social” in the sense that it gets better the more people use it, since more people can contribute more knowledge, or can correct details in existing knowledge for which they are experts.

“Social” software like wikis enable the creation of communication and relationships between individuals as well as between groups, and it supports the establishment, maintenance, and extension of social networks. We will see next that this concept of social software can be found elsewhere as well.

4.2 Social Networks

According to [13], social networks bring another dimension to the Web by going way beyond simple links between Web pages; they add links between *people* and *communities*. In such a network, direct links will typically point to our closest friends and colleagues, indirect links lead to the friends of a friend, and so on. In this section, we take a brief look at social networks and the impact they are having on the Web today.

The Networked Generation in 2006

Let us return to a picture we started drawing earlier in this section, and imagine an 18-year old today. This youngster not only grew up with the Web, he or she has never considered it to be a utopia, but has probably started using it a few years back already! Moreover, the younger generation is well underway with respect to abandoning traditional media. Recent investigations such as [19] show that the "networked generation" (young adults like our 18-year old) is driving a radical shift in media consumption. British broadband users spent on average 12.7 hours per week online. Seventy percent have used some kind of social networking site such as MySpace. Moreover, "there is also evidence of a significant difference in communications usage patterns between young adults and the general population: for example, 16-24 year olds [in 2006] spend on average 21 minutes more time online per week [than 2005], send 42 more SMS text messages, but spend over seven hours less time watching television." In other words, the 16 to 24 year olds are spurning television, radio, and newspapers in favour of online services. Increasingly households are turning to broadband and digital TV. As [19] states, "this generation has grown up with new technologies – and it is this generation for whom the uptake is instinctive." Yet the report also found technology adoption is not limited to the young. "The sector is being transformed by greater competition, falling prices and the erosion of traditional revenues and audiences."

Thus, while it initially took a while until average users and consumers started to accept the Web as a medium or platform, or actually "trusted" it, the young generation is now growing up with it and is hence integrating it into everyday life much more comprehensively. This observation is, by the way, perfectly in line with Friedman's Flattener No. 10, the steroids, and the way in which they have made their way into everyday life and usage.

Social Networking and MySpace.com

The information available on the Internet and the Web tools by which this information has meanwhile become accessible has led to the establishment of a number of distinct Internet *communities*, i.e., groups of people with common interests who interact through Internet and the Web. Today, we can identify at least the following types of communities:

- Communities of *transactions* which are characterized by the fact that they facilitate buying and selling as well as auctioning;
- Communities of *interest* which commonly center around a specific topic, e.g., diecast cars, health food, dietary supplements;

- Communities of *relations* which are organized around life experiences, e.g., traveling New Zealand on a budget, coping with breast cancer, or coming from the same high school;

- Communities of *fantasy* which are based on imaginary environments and game playing, e.g., SecondLife or Entropia.

The considerable change in perception and usage described above has opened the door for the present-day willingness of people to share all kinds of information, private or otherwise, on the Web, for a new "open culture" as observable in blogs, wikis, and, for example, at MySpace.com. MySpace was founded in July 2003 by Tom Anderson, its current president and CEO, Chris DeWolfe, and a team of programmers. Since MySpace is considered a "social networking" site, we have to say a few words about this term before we take a closer look at this service.

The Internet offers various ways to make contact with other people, including email, chat rooms, blogs and discussion boards (which, unfortunately, are also heavily exploited by spammers, hackers, and other users with unethical intentions). However, while these services just support ad-hoc interaction or focused discussions on a particular topic (or less focused conversations on the world at large), a *social network* goes a step further and is typically the result of employing some software that is intended to focus on building an online community for a specific purpose. Many social networking services are also blog hosting services where people can deliver longer statements than they would otherwise, and the distinction between blogs and social networks is often blurred. Social networks connect people with different interests, and these interests could relate to a specific hobby, a medical problem, an interest in some specific art or culture. Often, members initially connect to their friends which they know, for example, from school or college and later add new contacts, e.g., from their professional life, found through the Web.

While there are, as mentioned, obvious possibilities for misusing such a service, there are even professional business applications of it today. Indeed, consider a large, multi-national company comprising numerous departments, employees, and products. Such a company typically has a *knowledge management* problem, since there is a large body of knowledge available among its employees, yet making effective as well as efficient usage of that knowledge is not always easy. For example, a particular project might need expertise from a variety of distinct areas, and it may not be straightforward to spot the appropriate people throughout an enterprise who can actually contribute that expertise. Or a product might have a design flaw or a technical or production problem, and the best help would be to identify those people in the company who can find the cause of the flaw and eliminate it. Numerous techniques and systems have been developed to support knowledge management, to establish profiles of employees' expertise, and to bring the right people together for a particular job or project, yet many companies are still struggling with this issue and have no good policy of handling it properly.

A social network can act as a means of connecting employees of distinct expertise across departments and company branches and help them build profiles in an often much easier way, and it can do so in a much cheaper and more flexible way than traditional knowledge management systems. Once a profile has been set up and published within the network, others can search for people with particular knowledge and "connect" to them. A typical example of a social network often used professionally and for business purposes is LinkedIn, a network that connects businesses by industry, functions, geography and areas of interest

(http://www.linkedin.com); other such examples include XING and the *Open Business Club* (openBC, www.xing.com) as well as Ryze Business Networking (www.ryze.com). Since social networks can usually be set up free of charge (or for a low fee), they are an attractive opportunity for a company to create and expand their internal contact base. However, a social network needs not be restricted to a company's internal operation; it may as well include the customers to which the company sells goods or services (and hence be used for *customer relationship management*).

The social networking site that is probably most popular among members of the "networked generation" is certainly MySpace, a Web site offering an interactive, user-supplied network of friends, personal profiles, blogs, groups, photos, music, and videos. MySpace also features an internal search engine and an internal e-mail system. According to Alexa Internet, a company monitoring Internet traffic, it is currently the world's 4th most popular English-language website. The service has become an increasingly influential part of contemporary popular culture, especially in English speaking countries. Figure 21 shows a screenshot of the MySpace home page (taken on November 2006), which shows the *login* area, the *cool new videos* area, the *cool new people* area, as well as other areas related to music, videos, specials and the functionality of mail, blog, forum, groups, events, classifieds, etc.

A user profile at MySpace has two standard sections, "About Me" and "Who I'd Like to Meet," as well as several others, such as "Interests" and "Details." Fields in the latter sections will not be displayed if not filled in. Profiles also contain a blog with fields for content, emotion, and media. MySpace supports an uploading of images, one of which can be chosen the image to be seen on the main page of the profile and as the image that will appear next to the user's name on comments, messages, etc. Like others sites, MySpace now supports an uploading of videos as well. A core MySpace functionality that also makes it a "social" site is the *User's Friends Space*, which contains a count of a user's friends, a "Top Friends" area, and a link to view all of the user's friends. Each user can choose a limited number of friends to be displayed on his or her profile in the "Top Friends" area. In the "Comments" section a user's friends may leave comments for everybody to read, where each comment could require approval or even be deleted by the user. MySpace also allows users to modify their pages through an application of HTML and CSS and hence to create a *personalization* of their page. Users can put music and movies into their profiles, a feature intensively used, for example, by bands in order to promote their music.

More details on how to use MySpace can be found in [14]. To complete the picture, we also mention a few other sites where social networks can be created or set up: *Friendster* is seen today as the site that gave a big push to social networking since 2003, although the development itself just like blogs and wikis goes back to the 1990s (www.friendster.com). Friendster founder Jonathan Abrams, noticing that its popularity has been on the decline recently, has meanwhile founded *Socializr*, where published information is supposed to be more important than the community member who publishes it (www.socializr.com). Another site is *Ning*, which distinguishes groups from photo and video networks (www.ning.com). *Tribe* is a service that looks at the IP address of a user's machine and redirects her or him to a domestic site to find friends more locally (www.tribe.net). *Pendosoft* is a Swedish company offering social networking software, where a sophisticated client (and hence not the ad-hoc user) can retain some control over the code included in the network's site (www.pendosoft.com). More on the

entertainment side are places like the Japanese community entertainment network *Mixi* (www.mixi.com) or the *Singshot* network for online singers and Karaoke artists (www.singshot.com). MoveOn is a community of people (actually a roof for several organizations), "from carpenters to stay-at-home moms to business leaders," who are interested in political issues in The US (www.moveon.org). For the sake of completeness we mention that online communities actually go back to 1985, the founding year of *The Well* (www.well.com) which still exists today; another early community, almost as old as the Web itself, is *Bianca's Smut Shack* (http://www.bianca.com/shack/).

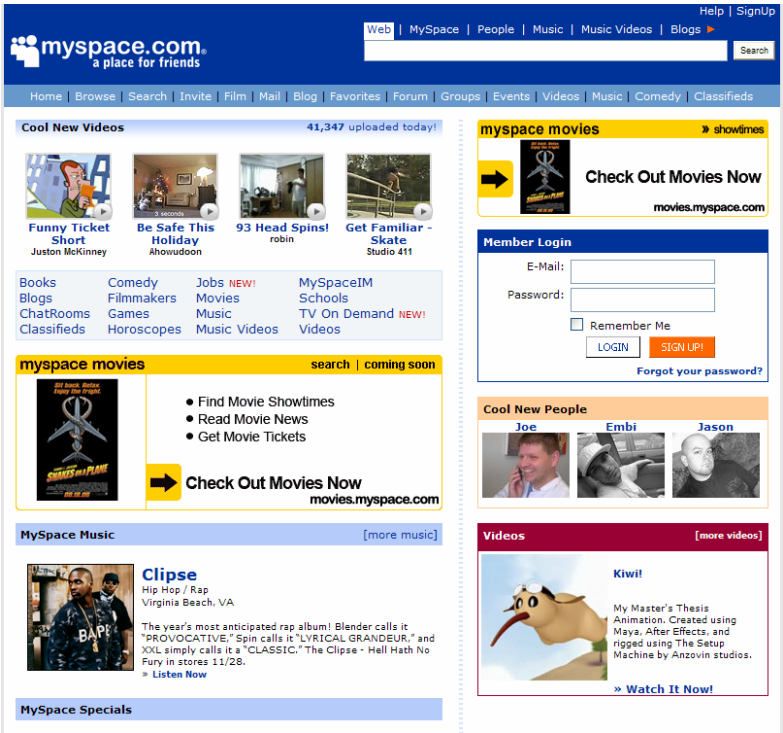

Figure 21: MySpace.com Home Page.

We finally mention YouTube in this context, which can be seen as a mixture of a video blogging site and a social networking site (www.youtube.com), but which exhibits an interesting phenomenon: Due to the size of networks nowadays existing on the Web, and due to the functionality which the sites mentioned above nowadays offer, IT and media companies have discovered their interest in social networks. To be on the safe side, they have started acquiring such sites, e.g., Google has bought YouTube, while Rupert Murdoch of News Corporation bought MySpace (see the cover story in *Wired* magazine, July 2006).

Social networks on the Web have also triggered a renewed interest in sociological questions regarding how the Internet and the Web are changing our social behavior (including the question of why people send out spam mails and try to hack into other people's accounts). Social networks can be extracted from the Web, for example, by looking at people's home pages or pages they have set up with one of the services mentioned above. Home pages typically tell about people's interests, hobbies, idols, maybe even family and friends, and they

link to corresponding other pages. An analysis of such a page in the style it is done by the indexer of a search engine can then reveal fragments of texts or individual phrases that are also used by other people on their home pages, and in this way commonalities, common interests, relationships, and even entire communities can be discovered.

As an example, The Stanford Daily reported the following on June 1, 2001 under the header "New Web site weaves social web": "A new Web site sketches the 'Stanford Social Web' by deducing the connections between people from the information contained on their Web sites. The Web site was developed by Lada Adamic, an applied physics graduate student at Stanford, and Eytan Adar, who works at the Hewlett-Packard Sand Hill Labs. Their research was supported by the Internet Ecologies Area of the Xerox Palo Alto Research Center. According to Adamic and Adar's research proposal, 'the links and text on a user's homepage and the mailing lists the user subscribes to are reflections of social interactions a user has in the real world.' Thus, by searching the Web sites of Stanford students for matching text and analyzing the links between Stanford students' pages, it is possible to recognize pairs of friends or acquaintances…. Provided that people's friends have Web sites hosted by a Stanford server, the Stanford Social Web can provide an intriguing sketch of their social lives. After entering a Stanford system user name onto the main Web site, one is led to a page showing a diagram of one's friends mapped out in the form of a constellation. The links between friends are inferred from the pattern of in-links and out-links that connect their Web sites. The Stanford Social Web not only indicates existing ties between students, but suggests pairs of students who share common experiences or interests, regardless of whether they are already acquainted with each other. The text on students' Web sites is compared within categories such as 'cities,' 'persons' and 'organizations,' in order to generate a 'likeness score.'"

On a broader scale, *social network analysis* investigates metrics that measure the characteristics of the relationships between the participants of a network. In particular, such an analysis looks for *ties* (between pairs of participants) and their *strength* as well as their specialization into *bridges*, for *triangles* (involving three people), and for issues such as the *clustering coefficient* of a participant or the *density* of the network. For a brief introduction, the reader is referred to [13], for a more detailed one to [5].

In conclusion, the current expectation for several years to come is that user participation in the Web, which already has increased considerably in recent years, will continue to grow, and that the social life of individuals, families, and larger communities of people will increasingly be enriched by online applications from social Web networks. It remains to be seen whether this will indeed have anything more than "local" effects in the sense that people may now take their network of friends online, but do not necessarily include people they have never met or that they most likely will never meet. On the other hand and as a result of this development, the social Web will become increasingly relevant to companies as a vehicle for marketing and communication.

5 Merging the Streams: The Arrival of "Web 2.0"

In this report we have tried to draw a picture of the history of the Web that is determined by three parallel streams of development: the *applications stream* that has brought along a number of services anybody can nowadays use on the Internet and the Web, the *technology stream* which has under-fed all of this with fast moving and comprehensive advances in networking and hardware technology and quite a bit of progress regarding software, and finally the *user perception and participation stream* (which we might also call the *socialization stream*) which has changed the way in which users, both private and professional ones, perceive the Web, interact with it, and publish their own information on it. These three streams have brought along a number of techniques, technologies and usage pattern that nowadays converge, and the result is what has received the preliminary and – as of 2006 – extremely fancy term "Web 2.0." We will next try to provide a brief characterization of what is behind the term, as well as in a summary of what has led to it. In summarizing the discussion from this report, we have seen the emergence of

- The arrival of HTML, of the browser, and of search as a dominant activity in many Internet applications these days;
- Electronic commerce as an addition to (and often a replacement of) traditional retail through stores, now being a major, if not the only source of income for many businesses, and also a major driver of new algorithmic applications of software (think, for example, of data mining through transaction data or automatic recommendation systems as made popular by Amazon.com and others);
- Advertising as a major business model and way of making money on the Internet, in particular as a vehicle that can provide the financial basis for services that are offered for free to end-users;
- The long tail as an illustration of a number of phenomena that the modern Web allows to respect, in particular advertising in niche markets once the necessary technology is available as a commodity (and, as a consequence, doing commerce in these niches);
- Customer feedback, blogging, and wikis as novel ways of obtaining user input;
- A disappearance of the general consensus that goods and services that are taken up basically have to be paid for, and a replacement of this consensus by the questionable understanding that electronic goods, which are transported mostly by being copied from one place to another, should more or less be for free; this is technologically supported by the advancement of file-sharing networks and psychologically by the open-access movement;
- HTML scripting, XML, and AJAX as fundamental language and development paradigms upon which present-day Web applications can be built;
- Service orientation and content syndication as new ways of providing software and content and of composing applications;

- Online communities, online encyclopedias, and social networks as (often anonymous) places where people can create collective knowledge and intelligence, and can meet and share experiences, made popular, for example, by the "MySpace generation."

When we compare these core aspects of the "history" of the Web that we have seen so far (where it may even be a bit far fetched to use the term "history" only 13 years after its inception), we can recognize that the Web has emerged from a medium where a few people centrally determined what all others had to use to one where very many people participate and jointly create and publish content. In other words, while initially content was mostly *read from* the Web, content is nowadays more and more *written to* the Web; this is why some people speak of Web 2.0 as the "read/write Web."

An immediate consequence of the fact that more and more people publish on the Web through blogs, wikis, communities, tagging, and otherwise is that increasing amounts of *data* have to be stored. More data arises from commerce sites, where each and every user or customer transaction leaves a trace in a database. As mentioned, search engines have already discovered this and started incorporating blogs into their crawling activities, and companies commonly employ data warehouse technology for online analytical processing or the application of data mining tools to large data collections in order to generate new knowledge; as a common saying in data mining goes, "there's gold in your data, but you can't see it." Making the most of large data collections will be even more crucial in the future than it has been in the past. This is especially true for data collections that are in some way unique or specialized; control over such a collection will be an invaluable asset in the future. However, it should be noted that this is not a *new* fact; indeed, holding control over a specific data pool has always been very valuable, the major difference today being that so many data collections are available online.

The technology stream has contributed a major paradigm shift in software development, namely a departure from large and monolithic software applications to light-weight services which ultimately can be composed and orchestrated into more powerful services that finally carry entire application scenarios. While the original Web service paradigm shown above in Figure 16 is not in full swing as far as the Web is concerned, a much simpler version is, which boils down to what we call "Web procedure call" (in reference to the long known "remote procedure call". Service-orientation especially in the form of service calls to an open application programming interface (API) that can be contacted over the Web as long as the correct input parameters are delivered have not only become very popular, but are also exploited these days in numerous ways, for the particular reason of giving users an increased level of functionality from a single source. A benefit of the service approach to software development has so far been the fact that platform development especially on the Web has received a high momentum in recent years. Yet it has also contributed to the fact that services which a provider delivers "behind the scenes" to some well-defined interface can be enhanced and modified and even permanently corrected and updated without the user even noticing, a phenomenon that has been called "perpetual beta" by Tim O'Reilly and others [18].

Taking the developments we have described together, considering them in conjunction, and putting them into a perspective of time and evolution, it becomes clear that the aspects we have described in our three streams above represent the core of the developments and the evolution of the Web, the Internet, its underlying technology, and the user perception of all this, so that it is justified to say that the essence of Web 2.0 boils down to the following three core aspects:

- Ways to utilize and combine *data* and *data streams* from various sources into a representation that allows for the derivation of new information or added value; the appropriate and intelligent (and legal) utilization of data that is kept current by its owner has become a driver of many new applications that nowadays come under the name "mash-ups;"

- *Functionality*- as well as service-oriented approaches to build new applications as a composition of other, already existing ones and in order to enrich user experiences on the Web, or to create "Rich Internet Applications" (RIAs), and the services provided let the Web exhibit features previously known from stationary computers only;

- Tagging, blogging, and "wiki-ing" as important steps into a *socialization* of the Web, where all of a sudden a user does no longer consider his personal entries private anymore, but makes them available to friends, a certain community, and ultimately the general public, and where this often even leads to an improvement of the underlying platform.

These core aspects can be seen as three dimensions of data, functionality, and socialization spanning a "Web 2.0 space" into which current Web developments can be positioned. Indeed, both Wikipedia and blogs can be seen as typical examples. A blogging site puts most emphasis on data (the data that the user puts into his or her blog) and on socialization (by letting other users read the blog entries and comment on them). Wikipedia, on the other hand, is a good example of a system with advanced functionality, yet also lots of data and an important social component based on expertise, social behavior, and trust.

We finally mention that the term "Web 2.0" is also under questioning already regarding its lifespan. For example, an article in MIT's *Technology Review* in December 2006 asked the question "What comes after Web 2.0?" and reminded its readers of several other developments which might be overlooked within the current Web 2.0 hype. The most prominent of them is the *Semantic Web*, a vision that has been laid out by Web inventor Tim Berners-Lee several years ago and whose goal is to add more "semantics" or meaning to the Web and its applications, so that it becomes easier for them to talk to each other. Many Web sites also discuss the term "Web 3.0" already and see a fusion coming of semantic technologies and architectures of participation as we currently experience them. On the more funny side, Jonathan Abrams, the aforementioned founder of Socializr, states on http://www.socializr.com/about that "Socializr is a 'Web 3.1' company. 'Web 3.1' is an arbitrary and silly label like 'Web 2.0', but even sillier, and 55% cooler! And everyone knows nothing works right until the 3.1 version." This should let us handle the term with care, and should prevent us from overestimating the novelty behind it.

References

[1] Alonso, G., F. Casati, H. Kuno, V. Machiraju (2004): *Web Services. Concepts, Architectures and Applications*. Springer Verlag, Berlin, Germany.

[2] Aspray, W., F. Mayadas, M.Y. Vardi (eds.) (2006): *Globalization and Offshoring of Software – A Report of the ACM Job Migration Task Force*; Association for Computing Machinery, New York.

[3] Anderson, Ch. (2006): *The Long Tail. Why the Future of Business is Selling Less of More*; Hachette Book Group, Lebanon, Indiana. See also the following Web page of Wired Magazine for a short article: http://www.wired.com/wired/archive/12.10/tail.html

[4] Battelle, J. (2005): *The Search – How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture*; Portfolio (Penguin Group), New York.

[5] Brandes, U., Th. Erlebach (eds.) (2005): *Network Analysis – Methodological Foundations*; Springer-Verlag, Berlin, LNCS 3418.

[6] Brin, S., L. Page (1998): The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks* 30, pp. 107-117.

[7] Denning, P., J. Horning, D. Parnas, L. Weinstein (2005): Wikipedia Risks; *Communcations of the ACM* 48 (12), December 2005, p. 152.

[8] Grandinetti, L. (ed.) (2006): *Grid Computing: The New Frontier of High Performance Computing* (Volume 14); Morgan Kaufmann Publishers, San Francisco, California.

[9] Han, J., M. Kamber (2006): *Data Mining: Concepts and Techniques*; Morgan Kaufmann Publishers, San Francisco, California, 2nd edition.

[10] Friedman, Th.L. (2005): *The World is Flat – A Brief History of the Twenty-First Century*; Farrar, Straus and Giroux, New York.

[11] Harold, E.R., W.S. Means (2004): *XML in a Nutshell*; O'Reilly Media, Sebastopol, California 3rd edition.

[12] Laudon, K., C.G. Traver (2007): E-Commerce: Business, Technology, Society; Prentice-Hall, New Jersey, 3rd edition.

[13] Levene, M. (2006): *An Introduction to Search Engines and Web Navigation*; Addison-Wesley, Harlow, U.K.

[14] Magid, L., A. Collier (2007): *MySpace Unraveled – A Parent's Guide to Teen Social Networking*; Peachpit Press, Berkeley, California.

[15] Miller, M. (2007): Google.pedia – The Ultimate Google Resource; Que Publishing, Indianapolis, Indiana.

[16] Musciano, C., B. Kennedy (2006): HTML & XHTML: The Definitive Guide; O'Reilly Media, Sebastopol, California, 6th edition.

[17] Norman, D.A. (1999): *The Invisible Computer: Why Good Products Can Fail, the Personal Computer Is So Complex, and Information Appliances Are the Solution*; The MIT Press, Boston, Massachusetts.

[18] O'Reilly, T. (2005): What Is Web 2.0 – Design Patterns and Business Models for the Next Generation of Software; http://www.oreilly.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html.

[19] Office of Communications (2006): 2006 Ofcom Report; Ofcom, London, downloadable from http://www.ofcom.org.uk/media/news/2006/08/nr_20060810.

[20] Olim, J., M. Olim, P. Kent (1999): *The CDnow Story: Rags to Riches on the Internet*; Top Floor Publishing.

[21] Payne, A. (2006): *Handbook of CRM: Achieving Excellence through Customer Management*; Butterworth-Heinemann Elsevier.

[22] Pogue, D., J.D. Biersdorfer (2006): *The Internet – The Missing Manual*; Pogue Press O'Reilly, Sebastopol, California.

[23] Pokorny, J. (2004): Web Searching and Information Retrieval; *IEEE Computing in Science & Engineering*, Vol. 6, No. 4, July/August 2004 issue, pp. 43-48.

[24] Roberts, M.M. (2006): Lessons for the Future Internet – Learning from the Past; *Educause Review*, July/August 2006 issue, pp. 16-24.

[25] Tanenbaum, A.S. (2003): Computer Networks; Prentice-Hall, Upper Saddle River, New Jersey, 4th edition.

[26] Tanenbaum, A.S, M. van Steen (2007): *Distributed Systems – Principles and Paradigms*; Prentice-Hall, Upper Saddle River, New Jersey, 2nd edition.

[27] Vise, D.A. (2005): *The Google Story – Inside the Hottest Business, Media and Technology Success of Our Time*; Macmillan, London.

[28] Vossen, G., St. Hagemann (2007): *Unleashing Web 2.0 – From Concepts To Creativity*; Morgan Kaufmann Publishers, San Francisco, CA, to appear.

[29] Yang, J. (2006): *The Rough Guide to Blogging – Navigate the Blogosphere*; Rough Guides Ltd., London.

Working Papers, ERCIS

Nr. 1 Becker, J.; Backhaus, K.; Grob, H. L.; Hoeren, T.; Klein, S.; Kuchen, H.; Müller-Funk, U.; Thonemann, U. W.; Vossen, G.: European Research Center for Information Systems (ERCIS). Gründungsveranstaltung Münster, 12. Oktober 2004. October 2004.

Nr. 2 Teubner, R. A.: The IT21 Checkup for IT Fitness: Experiences and Empirical Evidence from 4 Years of Evaluation Practice. March 2005.

Nr. 3 Teubner, R. A.; Mocker, M.: Strategic Information Planning – Insights from an Action Research Project in the Financial Services Industry. June 2005.

Nr. 4 Gottfried Vossen, Stephan Hagemann: From Version 1.0 to Version 2.0: A Brief History Of the Web. January 2007.