Search engine freedom: on the implications of the right to freedom of expression for the legal governance of Web search engines

van Hoboken, J.V.J.

Publication date
2012

Chapter 2: A Short History of Search Engines and Related Market Developments

2.1 The Internet, the Web and the rise of navigational media

2.1.1. Early visions of navigation in digitized information environments

The way in which digital computing would lead to a revolution in information and knowledge navigation was already being explored more than half a century ago, when computers were still a rarity and neither the Internet, nor the World Wide Web existed. Most famously, Vannevar Bush, in his article 'As We May Think', envisioned the 'memex', an electronic device in which individuals would store their books, records and communications on micro-film, and which would be consultable through a system of indexes and speedily navigation.[10] The users of the memex would be able to tie different pieces of knowledge together and compose their own trails in the body of information available on the memex. These ties and trails would remain available for later consultation and use.

Bush imagined this memex to help society overcome the limitations of the scientific organization of knowledge through the traditional indexing and storage of paper-bound information. His work was, and still is, part of the scientific literature relating to knowledge and libraries, a scientific field which was actively addressing the issue of how to organize the ever growing field of human knowledge through the use of new technologies. In the same field, but almost two decades later, Licklider continued this endeavor with a research project on the characteristics of the future library – to be precise, the library of 2000.[11] Licklider, mentioning Vannevar Bush as his main external influence, started with the same assumption: knowledge was growing at a speed beyond society's capacity to make use of it.[12] The primary reasons for this growing discrepancy, he claimed, were the limitations of paper-bound knowledge from the perspective of the user's need to retrieve relevant information. The only solution, according to Licklider's team, would be a fusion of the computer and the library into what they would end up calling a 'precognitive system'. Their work considered the feasibility of such a system on the one hand and the criteria it would have to fulfill on the other.[13] For instance, they concluded that the system would have to make the body of knowledge available when and where needed, foster the improvement of its organization through its use and converse and negotiate with the user when he or she formulates requests.

These early theoretical developments related to the use of digital technology to consult digital collections of information gave birth to the field of information retrieval, the science or field of information engineering relating to the search and retrieval of electronic materials and of the information within such materials.[14] The scientific roots of current Web search engines lie in this field of information retrieval.[15] But already more than 50 years ago, many seemingly obvious but fundamental improvements were conceptualized and tested for information retrieval more generally, that realized

[10] Bush 1945. For a discussing see Stefik 1996, pp. 15-23. Paul Otlet may have been the first to explore these issues. See Wright 2008; Rayward 1990.
[11] See Licklider 1969.
[12] See Stefik 1996, p. 24.
[13] See Licklider 1969, pp. 32-39.
[14] See Singhal 2001; Lesk 1995.
[15] See Röhle 2010, p. 17.

the promise of the computer not only as a basic storage unit for information and also of making this information more easily accessible. In the 1950s, for instance, information scientists proposed the use of statistical text measures for the relevance of documents. Maron and Kuhns conceptualized the use of words as indexing units for documents and the measuring of word overlap, i.e. the similarity of the entered search query to the set of indexed words in available documents, as a criterion for retrieval relevance.[16] These and related ideas caused a paradigm shift in thinking about the problem of information retrieval: relevance in information retrieval systems would no longer be a binary affair, meaning that a document was simply relevant or not, but would become a prediction of how valuable a document would be for a user of the system. This prediction would be based on an inference of the searcher's input and the contents of the documents in the system.[17] The field of information retrieval has made rapid progress ever since and has made a major contribution to the conceptualization and development of the later search engines for the Internet and the World Wide Web, which this study is focusing on.[18]

The following section will shortly explore the historical societal context of search engines by looking at the issue of findability and the state of search technology from the start of the Internet to the current public networked information environment. Notably, a choice has been made to focus on the historical background of Web search engines from the perspective of end-users and to see them as the current end-product of the development of instruments for effective navigation and retrieval which evolved together with the expanding digital information environment.

2.1.2 The Internet: connecting the nodes

When the Internet, or to be more precise the 'ARPANET', was developed in the end of the 1960s,[19] the network was more about computer resource sharing than about the sharing of knowledge and information. In that sense it was far from the visions of Bush and Licklider about the use of networked computing to create a memex or the future library. Notwithstanding this primary purpose of sharing computing resources, the issue of findability, i.e. users of the network being able to know what computer resources, documents or other users were available on or connected to the network, was an important one that had to be addressed. It was partly resolved by the funding through ARPA of a Network Information Center (NIC) at the Stanford Research Institute (SRI). The NIC was created by the SRI research group led by Douglas Engelbart, a pioneer in human computer interaction and networked computing. The NIC maintained several directories essential to the use of the network.[20]

Because of their research into the opportunities of better handling of digital resources, Engelbart's research group was an attractive candidate to fulfill the role as envisioned for the Network Information Center. Nonetheless, the NIC was not particularly successful in the task of overseeing the resources that

[16] Maron & Kuhns 1960, pp. 216–224. For a discussion of Kuhns work see Singhal 2001. For a discussion of Luhn's propositions regarding text-statistical measures as a model for relevance, see Röhle 2010, pp. 113-118.
[17] See Maron 2008, pp. 971–972.
[18] See Singhal 2001. See also Baeza-Yates & Ribeiro-Neto 2011.
[19] For a history of the Internet, see Abbate 2000. See also Hafner & Lyon 1996, Stefik 1996.
[20] Abbate 2000, p. 59.

were available on the ARPANET. It proved hard to organize a reliable and complete directory of network resources and capabilities. [21] Ultimately, a lot of information about the network was shared informally or off the network. When the electronic mail protocol was introduced, the newsletter became an important new means of sharing resource information. The *ARPANET News*, for instance, had a special section featuring certain network resources.[22]

The current Internet grew out of the ARPANET of the 1970s, but besides ARPANET, there were many other computer networks that offered similar possibilities, such as the global network based on the X25 network standard that was widely deployed by the telecommunications industry in the 1980s. Ultimately, various standardization efforts, including the introduction and promotion of the TCP/IP standards in the 1980s, and the switchover of other networks to this defining Internet standard, helped to create the global network of networks, the Internet, that we know today.[23]

Public access to the network remained limited until the beginning of the 1990s. Throughout the 1980s several private networks provided services to meet the popular demand that was shaped by the early personal computer revolution of that time. Dial-in networks such as CompuServe, AOL, and Prodigy, and a variety of smaller Bulletin Board Services (BBSs; accessible for computer users by calling in over regular phone lines), offered the possibility to access information and entertainment, post messages and play early network-based computer games. These BBSs were very popular in the beginning of the 1990s and many of the early legal issues related to the Internet involved BBSs.[24]

Notably, the ARPANET was not a public resource; access to the network was restricted. This remained the case until the Internet became publicly accessible in the 1990s. Consequently, the resources that were available on the network were not part of the public information environment either. And those who had access to the network could, in principal, not freely access all the information on the net, unless they had (implied) permission to do so. Clearly, these features of the ARPANET had important implications for the state of findability on the network, as not all the material on the network was freely accessible to all the users, let alone potential directory and or search engine providers. This excerpt from set of guidelines about the use of ARPANET from the Defense Communications Agency (DCA) in an ARPANET Newsletter from 1981 shows the restrictions on the accessibility and further use of information on the ARPANET:

> "*Files should not be FTPed by anyone unless they are files that have been announced as ARPANET-public or unless permission has been obtained from the owner. Public files on the ARPANET are not to be considered public files outside of the ARPANET, and should not be*

[21] See Abbate 2000, pp. 85-89, 213.

[22] See Hafner & Lyon 1996, pp. 229-230. ARPANET News was edited by the same Stanford Research Institute.

[23] Abbate 2000. On the general engineering aspects of computer networking, see Kurose & Ross 2009.

[24] Examples are the raid of the Private Sector BBS of Hacker Magazine 2600. See '2600' 1985. For a discussion of the German CompuServe battle against access for German citizens to illegal content which started in 1995 (Germany), see Determann 1999. See also Playboy Enterprises Inc. v. Frena, 839 F. Supp. 1552 (M.D. Fla. 1993); Religious Technology Center v. Netcom On-Line Communication Services, Inc., 907 F. Supp. 1361 (N.D. Cal. 1995).

transferred, or their contents given or sold to the general public without permission of DCA or the ARPANET sponsors."[25]

In addition, even though users of the early Internet of the 1970s and 80s may have been able to access material hosted elsewhere, before the introduction of browsers and the World Wide Web, it was relatively hard to find (unknown) content on the Internet. There were no search engines yet and files in remote locations were typically accessed by using the file transfer protocol (FTP) protocol. This protocol was designed to transfer files over the network and not to find them effectively. The DNS system, which added a human understandable address space to the numerical Internet address space did help Internet users to remember the locations of known organizations and hosts, but its value from the perspective of effective navigation online, particularly in view of the potential of information retrieval in digital information collections, was (and remains) limited.[26]

As the amount of resources available on the network grew steadily, effective retrieval became more and more of an issue, and specialized services were developed to keep track of resources and provide network users means to find materials. The first Internet search engine, Archie, was the first service to provide a searchable index of the titles of files available on anonymous[27] FTP servers on the network. It was developed by McGill University students in Montreal in 1990.[28]

As mentioned above, the FTP protocol had its limitations due to its focus on transferring materials over the network. These limitations of the Internet around 1990, in terms of the organization of content on the network to allow for the effective retrieval of material and its effective dissemination more generally, spurred the development and implementation of systems of additional protocols relating to the publication and organization of information on the Internet.[29] One of these sets of protocols was Gopher, which entailed a different way of organizing electronic materials on host sites on the network and ways of communicating with them from remote locations. The other and more famous one was the World Wide Web hypertext system, which will be discussed in the next section.

The Gopher system, which was released in 1991, relied on directory-based hierarchies for the storage and retrieval of information on the Internet.[30] In a Gopher environment, Internet users would be presented with directories of content available on the network much like they were used to in the typical text-based computer interfaces at the time.[31]

[25] Haugney 1981. The Defense Communications Agency was responsible for the legal management of ARPANET and with its guidelines l addressed the typical variety of legal issues related to the opportunities that the network offered to access and distribute information.

[26] For a more recent discussion of the navigational value of domain names, see Committee on Internet Navigation 2005.

[27] Anonymous ftp was commonly used to make material generally accessible for remote users, while not requiring users to log into the host server.

[28] See Sonnenreich & Macinta 1998, pp. 1-2.

[29] See Schatz & Harding 1994.

[30] Anklesaria et al. 1993.

[31] Media theorist Florian Cramer offers his website in Gopherspace. See Cramer, http://cramer.pleintekst.nl:70/.

The Gopher system for the publication of information on the Internet is of special interest from the perspective of the history of search engines because it had search and the effective retrieval of material designed into the system. Gopher involved so-called structured directory formats, calling for the organizing of online material in tree-like hierarchies, and moreover its design included the possibility of special full-text search servers which would help users of the network locate documents in specific domains. A second early search engine, Veronica, developed in 1992 at the University of Reno, focused on this new Gopher protocol and provided a directory of the hierarchies of Gopher servers available on the Internet. Like Archie, Veronica was limited to titles and did not offer full-text search.[32]

Notably, the specific focus on search and the organization of material on the network in the Gopher protocols was absent in the hypertext environment as introduced with the World Wide Web. Or maybe it is better to say it was deliberately left open. The World Wide Web, unlike Gopher, revolutionized the way in which the Internet was used as a public information environment. In combination with newly introduced browser technology, it marked a new phase in the use of the Internet for the sharing of information and ideas. And it also signified the real kick-off of the development of search engine services and technology.

2.1.3 The World Wide Web: Browsers, hyperlinks and spiders

The World Wide Web hypertext system was developed by Tim Berners-Lee and his colleagues at CERN (the European Organization for Nuclear Research), as a new way to organize information on the Internet. Building on existing ideas about hypertext and the memex vision of Vanevar Bush, the proposal for a World Wide Web in 1990 aimed to make online information more easily accessible in a universal format that would potentially link all online information together as a network of hypermedia nodes. As the first proposal for the World Wide Web stated:

> *The current incompatibilities of the platforms and tools make it impossible to access existing information through a common interface, leading to waste of time, frustration and obsolete answers to simple data lookup. There is a potential large benefit from the integration of a variety of systems in a way which allows a user to follow links pointing from one piece of information to another one. This forming of a web of information nodes rather than a hierarchical tree or an ordered list is the basic concept behind HyperText.*[33]

In other words, the World Wide Web hypertext system offered network users the opportunity to organize online information themselves by linking it together, instead of relying on more rigid hierarchical tree-structures such as in the Gopher system. Any contributor to online information, when using the hypertext markup language (HTML), would be able to link to any other available hypertext online resource, thereby integrating the new material with the rest in a universal 'web' of online materials. Network users would access the online environment with 'browsers', which would interpret the hypertext world and allow users to navigate it by going from node to node across the hypertext structured material on the Web. The nodes were to be identified by Uniform Resource Locators (URL),

[32] See Sonnenreich & Macinta 1998, pp. 2-3.
[33] Berners-Lee & Cailliau 1990.

which were based on the Internet host name space (DNS) and which would provide for a World Wide Web address space.

The World Wide Web hypertext system proved an enormous success and is generally seen as the application of the Internet that made it attractive to the masses.[34] Developed and first implemented in the global community of high-energy physicists, its use grew very rapidly after the release of Mosaic, an early publicly-available graphical Web browser that had more advanced capabilities such as allowing images to be shown as part of a Web page. Soon after, other commercial browsers became available on the market, such as the commercial Mosaic release called Netscape.

The organization of online information through a dynamic web of hyperlinked nodes, instead of a preconceived hierarchical structure, implies much more freedom for both users and contributors of the network than the hierarchical organization of materials in the Gopher system. The World Wide Web places emphasis on the ability of end-users to navigate online material effectively and relies on the knowledge of users - of all sorts - about the network and on their resourcefulness to provide the links to other available valuable materials on the network. Hence, this initial lack of organization of the hypertext environment implied an enormous opportunity for users to help 'organize' the World Wide Web and the navigation of information and ideas it made potentially possible. On the one hand, the World Wide Web design implicitly assumed that end-users and third parties would actually organize the Web. On the other hand, the demand for this organizing activity inclined steeply as the Web started to grow: there was more and more demand for 'useful link' web pages, directories and search services which would help users to find material located elsewhere.

The first Web search engines which responded to the demand for organized findability were, like Archie and Veronica, developed in the scientific community. The first crawler-based search engine, a search engine that uses a piece of software called a 'crawler' to access, analyze and index the World Wide Web automatically by following links for page to page, was the World Wide Web Wanderer, developed at MIT in the early 1990s.[35] Its main purpose was to analyze and report on the growth of the Web. The Wanderer automatically looked on the Web for available material and systematically stored data about this material, including its location in a central index which was called the Wandex.

Not everyone on the network welcomed the arrival of crawlers, also called bots or spiders, which automatically navigated the network to analyze its content. The network load caused by the Wanderer or similar software by repeatedly looking up material online to refresh their indexes, soon led to complaints and discussions about the ethical use of and proper restrictions on the deployment of crawlers. Notably, this discussion did not result in a ban on crawling activity, assuming that such a ban would be possible or enforceable. It did spur the development of an unofficial industry standard that allowed website hosts to give instructions to the crawlers indexing their sites. This *robots.txt* de facto standard, which is still generally followed today, was developed by Martijn Koster along with an index of the World Wide Web called ALIWEB. Instead of crawling the Web, ALIWEB relied on webmasters to

[34] See Abbate 2000, pp. 216-218. See also Berners-Lee 2000.

[35] See Sonnenreich & Macinta 1998, p. 3.

create and submit a special indexing file outlining the material they were publishing.[36] ALIWEB's 'anti-spider' model did not succeed in mobilizing webmasters enough to be able to create an index large enough to compete with other search engines. Instead, the crawler-based search engine model, in which the service would simply look itself for the available material, was actively pursued and soon a growing number of spiders was crawling the Web.[37]

In terms of the model of how to create an index of online material necessary to provide useful Web search services for end-users, the main rivaling model for the crawler-based search engine was the human-edited directory, of which the Virtual Library, Yahoo!, Looksmart and Magellan were all examples. Besides the Wandex, important early examples of crawler-based search engines were Excite and WebCrawler, WebCrawler being the first to index the complete documents on the Web and the first to provide a full-text search capability. Yahoo!, the most popular directory in the World Wide Web's history, grew out of a manually organized set of hyperlinks created by two Stanford students.[38] When it became more and more popular, they made their index of hyperlinks searchable.

Over the years the directory-based model for offering organized findability has slowly declined and the crawler-based model has become the standard for general purpose search engine services.[39] However, even today, the leading crawler-based search engine, Google, does still offer a directory as a part of its offerings to its users. In addition, some of their operations with regard to their crawler-based service increasingly rely on other directories,[40] or the kind of human judgment and intervention with regard to the relevance of online material which could be seen as a principal characteristic of human-edited directory-based services. As a result, the traditional distinction between crawler-based and human-edited directories has been blurred over the years.

2.2 The Web search engine

2.2.1. Web search engines: the birth of an industry: 1993-1998

Like in the case of browsers, the market soon picked up in the field of search engines, and commercial search engine providers and directories have been dominant ever since. As a result of the commercial nature of search engine services, the further development of search engine services and the innovations in this field are to a considerable extent a matter of business innovation, rather than only innovation in the scientific or technical sense of the word. Early search engine developers explored the various business opportunities related to Web search engines, developed new advertising models or licensing schemes, and explored strategic alliances with media conglomerates, the telecommunications and the ICT industry. Despite such activity, fundamental improvements to Web search technology are continuing to be made. Dominant Web search engines like Google heavily rely on cutting edge research in the fields

[36] Sonnenreich & Macinta 1998, p. 4.

[37] The idea that the different crawlers may be duplicating the same effort spurred the project Commoncrawl, a project to "maintain and make widely available a comprehensive crawl of the Internet for the purpose of enabling a new wave of innovation, education and research." See http://www.commoncrawl.org.

[38] See Sonnenreich & Macinta 1998, pp. 6-7.

[39] See Rogers 2009b.

[40] Crawler-based search engines can also use existing directories to help them to organize their index and rank references.

of computer science and electrical engineering, language processing, and network economics, and have also themselves been at the forefront of fundamental improvements in Internet and Web service engineering. For the purposes of this chapter, we will provide a general overview of the main developments in the search engine industry and the various business models that were invented and pursued by search engine entrepreneurs. The business model of search engines will not be analyzed in detail as a goal in itself.

Soon after the first web search services developed by researchers in the academic realm gained visibility, some of them acquired venture capital and went commercial. This is the first phase of the Web search industry, which political economist Van Couvering in her research on search engine bias, denotes as the phase of 'technical entrepreneurship'.[41] The dominant crawler-based search engine in this period, ranging roughly from 1993 until 1998, was AltaVista. Yahoo! was the most important directory online. Interestingly, AltaVista provided Yahoo! with crawler-based organic search results, complementing Yahoo!'s directory. Competition between different Web search services mostly focused on the size of the index - or directory - and the speed of response to user queries.

In this first stage, the business case for early search engines, like for many new online services, wasn't clear. The most common revenue stream was advertising, which on the early Web typically involved the placement of advertisements in the form of banners on a cost-per-view basis. Search engines and directories were attractive real estate for the placement of such advertisements as they attracted large numbers of Internet users. But apart from advertising, which linked search engines to the media industry, search engine providers started to rely on licensing, a long-established business model for software and related technology. By licensing their search engine software to destination websites or other services with high traffic, such as America Online (AOL) or Netscape, search engine technology companies could increase their distribution and secure revenue. These and other types of distribution deals became and remain an important field of competition between different search engine providers. These revenue sources were important because subscription-based business models, such as introduced by Infoseek in 1995, proved unsuccessful in the face of free services of comparable quality.[42]

Of special interest in the first stage of development of the search engine industry is the advertisement-based business model that was developed in late 1997 by business entrepreneur Bill Gross and implemented in the service GoTo.com.[43] Instead of crawling the World Wide Web, GoTo.com relied on the auctioning of keywords to the highest bidding online information provider. These bidders would not have to pay-per-view of their advertisement, which was the common way to sell advertisement space and is typically denoted by CPM (Cost per mille). Instead, bidders would only pay if a user would actually follow the advertised link to the bidder's site, a model denoted by CPC (Cost per click). This resulted in a shift in monetization of audiences to the monetization of actual traffic to destinations. GoTo.com was very successful and pioneered the syndication of paid search listings; around the year 2000 it had become the industry leader in the paid search market.

[41] See Van Couvering 2009.

[42] See Van Couvering 2009. See also Infoseek 1995.

[43] For a discussion, see Battelle 2006, pp. 101-104.

Table 2.2: a selection of notable Web search engines	
AltaVista	Crawler-based search engine (1995), market leader around 1996-1997, under ownership of Compaq (1998), CMGI (2000), Overture (2003) and Yahoo! (2003).
Archie	First Internet search engine (1990), provided a searchable index of titles of online resources.
Ask	Formerly known as Ask Jeeves (1996), initially modelled around concept of answering everyday questions of users. It was renamed Ask.com in 2005, is currently owned by IAC. It is said to have stopped producing its own organic results.
BING (Microsoft)	General purpose search engine service, formerly named MSN and Live.com, in which Microsoft invested billions of dollars to be able to compete with Google.
Blekko	New general purpose Web search engine, developed in California, went in closed alpha since July 2010, and has become publicly available in 2011.
Exalead	French search technology company (2004), participated in the Quaero project, is mostly focused on enterprise search.
Excite	Early crawler-based search engine which went the portal route with its merger with @Home in 1999.
Google	Web search provider (1998) coming out of Stanford; current market leader in the Web (search) services industry; made important improvements to the Web search experience for end-users since the end of the 1990s and implemented very successful paid listings program for search listings and the Web more generally.
Ilse	Early crawler-based search engine in The Netherlands, which stopped producing its own Web index, and is now owned by Sanoma.
Inktomi (HotBot)	Early crawler-based search engine software company coming out of UC Berkeley (1995), implemented into HotBot service which wasU.S.market leader in the late 1990s. Acquired in 2002 by Yahoo!.
Lycos	Popular portal in the end of the 1990s, seperate companies forU.S.and Europe (owned by Bertellsmann and Telefonica). In Europe Lycos portal included Web search Fireball and news search Paperball.
MetaCrawler	The first *meta* search engine (1995), searching various genuine search engines simultaneously and presenting those results to its users.
Open Directory Project	An open content volunteer-edited Web directory (1998), also known as dmoz, owned by Netscape (Oct 1998), which was in turn acquired by AOL (Nov 1998).
Overture (GoTo.com)	Founded by Bill Gross, pioneer of pay per click (CPC) and auctioning model and paid listings syndication.
Quaero	Politically inspired Franco-German search engine project that turned into two separate R&D industry investment programs for search technology in the broader sense.
Veronica	Early Internet search engine based on the Gopher protocol (1992).
Yahoo!	Early popular and commercially successful Web directory (1994), coming out of Stanford, branched off into various other personal services and advertisement products and acquired several Web search engines through its history. Stopped producing its own organic search results in 2010.
WebCrawler	Early Web search engine.
World Wide Web Wanderer	First crawler based Web search engine.

2.2.2 The birth of Google

Van Couvering lets the period of technical entrepreneurship end shortly after the first public offerings, amongst which the Yahoo!'s IPO in 1996 was one of the most significant ones in the Internet industry. The period that starts after that is a period in which one sees a tendency towards vertical integration

and a focus on the creation of, and partnerships in, so-called 'portals'. Most search engines companies in this period had a directory or a search engine at their core, but became focused on the presentation of all sorts of featured content and other services to their visitors. As Van Couvering shows, the featured content partnerships were seen as a way in which businesses could attract audiences to their content and services by paying these steadily growing portals for prominence. These vertical partnerships culminated in the vision of the 'fully-integrated portal' of the late 1990s, which amongst other things promised renewed control of the user's online experience for media conglomerates and telecommunications providers.[44] In line with the related tendency to vertical integration, several major deals were made that involved early search engine companies, such as the deals in 1999 between Infoseek and Disney and between Excite and @Home, which also involved AT&T.

Notably, as a result of the creation of portals and vertical partnerships involving featured content and services, the search engine was slowly downgraded in importance from being the core business to just a requirement or even an impediment to the portal's business model.[45] While the fully-integrated portal's focus was on keeping the user on the portal's sites, Web search engines in the strict sense tended to direct users away to other destinations online.

In hindsight, these developments opened up the space for Google to start its remarkable rise to dominance in the search engine industry. Google, like the early search engines, was developed in the academic realm, by computer science doctoral candidates Sergey Brin and Lawrence Page that were focusing on information retrieval science.[46] Google started as an experiment with a new ranking algorithm (PageRank) based on the network topology of hyperlinks on the World Wide Web.[47] PageRank was a global relevancy measure that assigned relevance to a document based on the weighted sum of incoming links to that document. The weight of the each link was determined by the relevance of that document itself and the amount of other links from that document.[48] Initially, the Google search service was clearly focused on providing the best search results possible, instead of seeing search engines as a means to a business end. Apart from venture capital and some important first distribution deals, for instance its deal with Netscape, the early Google did not have an advertisement-based business model and also no partnerships that involved featured content on their site. Instead, Google offered search results only, with a remarkably clean user interface that in no way resembled the cluttered portals and directories which were so common at that time.

By the time Google was introduced, existing search engines also increasingly suffered from third party manipulation of their relevance and selection criteria, and innovations and better business practices in this field were badly needed from the perspective of Internet users.[49] The typical selection and ranking of entries in the index in response to user queries by early search engines such as AltaVista proceeded in

[44] See Van Couvering 2009. See also, Meisel & Sullivan 2000.

[45] See Ince 2000. See also Edwards 2011.

[46] For a history of Google, see Vise & Malseed 2005; Battelle 2006; Levy 2011, Edwards 2011, Röhle 2010.

[47] Brin & Page 1998.

[48] See Brin & Page 1998. See also Langville & Meyer 2006.

[49] This manipulation of search results also soon led to various legal issues. See Nathenson 1998. For a discussion, see Chapter 8 and Chapter 10.

two steps. First, the query led to a subset of documents in the index that contained at least one of the query terms. This incentivized the use of irrelevant terms on websites in order to reach larger audiences in search engines. Second, the subset would be ranked according to basic information retrieval measures such as the amount of times that certain terms appeared on the website, the URL, and different hypertext meta-data, such as the field for the description of the content of the website. In response, other search engine optimization techniques were developed, such as various uses of meta-tags which would ensure better ranking in search results. As a consequence, the overall quality of search engines for users declined, whereas the need for effective navigational media grew alongside the rapidly growing World Wide Web.

Google's PageRank algorithm, which relied on a global measure for the relevance of websites, was in many ways motivated as a response to the growing infoglut and the manipulation of search engine results.[50] And in those early years after its launch, Google's focus on the quality of the search experience for its users gave it a competitive advantage. In 2000, after having displaced one of its main competitors, Inktomi, as the source of organic search results at Yahoo!, Google founder Larry Page was confident enough to state Google's superiority in terms of the relevance of Google's search results:

> "*We have very complex software that constantly analyses search results and can adapt itself to provide users with web pages that are more relevant to their search than from any other search engine.*"[51]

At the same time, the operational costs of general purpose search engines were steadily growing. Around the year 2000 the Web was estimated to already consist of more than 1 billion indexable pages. The crawling, indexing and speedy response to user queries on this scale demanded more and more fundamental innovations, knowledge and financial investment from search engine providers. In addition to its focus on improved ranking algorithms and the clean user interface, the success of Google to address these demands can help to explain its remarkable rise as the dominant search engine at the beginning of the 21st century.[52]

Over the years Google let go of their initial objection to an advertisement-based business model. It introduced the 'self-service ad program' Adwords in October 2000.[53] Since then, Adwords has been improved and perfected. Notably, in February 2002 the initial pay-per-view model was replaced by a pay per-click-model similar to the one used by Overture, the former GoTo.com. In May 2002, Google took over industry leader Overture's most important customer, AOL, with a major distribution deal which paved the way for Google's dominance in paid search listings.[54] Together with the extension of Adwords into the realm of general web publishing, i.e. the contextual advertising service AdSense, it solidified tremendous revenue streams for the company that it has used to finance additional free services for

[50] See Brin & Page 1998.

[51] See Foremski 2000.

[52] For accounts of the rise to dominance of Google, see Battelle 2006; Vise & Malseed 2005; Olsthoorn 2010; Vaidhyanathan 2011; Auletta 2009; Levy 2011; Edwards 2011.

[53] See Battelle 2006, pp. 125-127.

[54] See Gallagher 2002.

end-users, research and development and a range of acquisitions. These acquisitions include video platform YouTube in 2006, the online display advertising network DoubleClick in 2007, and the recent acquisition of ITA, a dominant search software and technology company in the travel industry.

2.2.3 Consolidation of the Web search industry: 2000-2011

Over the first decade of the 21st century the search industry has gradually consolidated further and only a few global market players dominate the market for general purpose search results in European countries and the Americas. This current period in the history of the search engine industry, is the one Van Couvering denotes with 'syndication and consolidation'.[55]

Consolidation has taken place on a number of levels.[56] First, many independent search engine providers were bought by other companies. The bursting of the dot-com bubble contributed to some of these acquisitions. Around the year 2000, Yahoo, for example, bought the search engine companies Overture, Inktomi, and AlltheWeb. At the time Yahoo! established ownership of Overture, Overture had already acquired AltaVista.

Second, many crawler-based search engines made the decision to stop producing search results themselves and enter into syndication deals with dedicated search engine providers instead. This meant that the amount of search result producers has declined correspondingly. Google proved particularly successful in establishing syndication deals, both for its organic results as well as for its paid listings.[57] These deals effectively secured access to the majority of Internet users for Google.

Third, general purpose search engines services started to offer more and more specialized services for their users, which implied that a simple Web search engine became less and less sustainable as a stand-alone business. Again, Google is best used as an example in this regard. It started to introduce more and more language specific services, it introduced image search in 2001, news search and product search in 2002, book search in 2003 (Google Print), geographic search in 2004 (Google Local), and ultimately the fully-integrated 'universal search' service in 2007. Many of these new features and services were made possible by the acquisition of smaller companies which had developed successful technologies to enabling these specialized services. Whereas many of these new additions could be seen as extensions of finding information, dominant providers in the Web search industry also started to offer different kinds of services to Internet users and thereby compete in other markets. Google, for instance, now offers a web-based email service (Gmail), an operating system for mobile devices (Android), and a cloud-based solution for document creation (Google Docs).

[55] See Van Couvering 2009.

[56] For figures, see Van Couvering 2009.

[57] For a discussion, see Levy 2011; Edwards 2011.

Table 2.1: Current Search Engine Market Share in terms of Query Volume[58]					
	Google	**Bing**	**Yahoo!**	**Ask**	**Other[59]**
The Netherlands (June 2010)[60]	94 %	1 %	1 %	0 %	4 %
Germany (18 May 2011)[61]	89%	4 %	2 %	1 %	4 %
France (March 2011)[62]	92%	4 %	1 %	0 %	3 %
United Kingdom (14 May 2011)[63]	90%	4 %	3 %	1 %	2 %
United States	65%	14 %	16 %	3 %	1 %

For the dominant search engine providers today search is only part of their business, but it remains one of the most important drivers in the industry. This may be illustrated most clearly by the decision of Microsoft to invest billions of dollars into the development of a search engine with their own organic and paid listings. Microsoft's MSN portal used to deliver search results of other search engine providers, including Google, but in 2011, after an investment of billions of dollars and two changes in names, Microsoft's Bing is now the second search engine in the western world as measured in search query volume. In fact, Microsoft has replaced Yahoo! as Google's main competitor in the web search industry, since Yahoo! has given up the competition in the field of organic search results. More specifically, in 2009 Yahoo! and Microsoft entered into a partnership which ended Yahoo!'s production of search results after a deal between Google and Yahoo! fell through because of the alleged anticompetitive nature as a result of Google's already dominant market position.

2.2.4 The Web search industry in Europe

If we look more closely at the search industry in Europe, the most important development has been the almost complete extinction of European-based search engines in terms of market share. Historically, in Europe similar phases of development can be found as described above, but European search engines have never successfully competed with the Web search giants from the Unites States. Since the second half of the 1990s, one can find a range of early Internet entrepreneurs in European countries that

[58] These figures are derived from different public reports of net statistic providers and are meant to serve as an indication of market share.

[59] In The Netherlands, meta-search engine Vinden.nl (3%), and Ilse.nl (1%); In Germany, T-Online (owned by T-Mobile but search results by Google) (2%), and others (2%); for France, SFR (1%) Orange (1%) and non-specified (1%); for the United Kingdom, non specified (2%); for the United States, AOL (1%).

[60] See Checkit.nl 2010.

[61] See Webhits 2011.

[62] See AT Internet 2011.

[63] See Hitwise 2011.

started local search and directory businesses.[64] But ultimately, most popular European crawler-based search engines stopped producing their own search results, went bankrupt, or only remained as brand destinations and domain-names in the hands of other companies.

In the Netherlands, for example, the crawler-based search engine Ilse had significant market share for some years. This service was launched back in 1996 by computer science student Wiebe Weikamp with two of his friends and was primarily focused on search results of particular interest for the Dutch population. In 2000 it was acquired by a large media company (VNU Uitgevers), and it stopped producing its own search results under the subsequent ownership of media company Sanoma. In Germany, Fireball and Web.de were strong local competitors. Web.de no longer produces its own search results, whereas Fireball shut down in 2002 after having been integrated with Lycos Europe in 2000. Lycos Europe's assets, after having been acquired by Telefonica and Bertelsmann and after having seen its search engine market share decline sharply since the end of the 1990s, were separately offered for sale in 2008. Fast, a successful Norwegian technology company with a strong search technology portfolio, was bought by Microsoft in 2008. [65] Exalead, a search engine technology company founded in 2004 in France, still exists as an independent European crawler-based search engine but is largely focusing on enterprise search services and business information management solutions. A notable example of a Web search engine that remains competitive at the national level is Yandex in Russia.

The dominance of United States companies in the sphere of the organization of information and ideas has not gone unnoticed and continues to spur political activity at the highest levels.[66] The most famous example of a European counter-initiative is Quaero, the European search engine project that never actually materialized into a service. Quaero was announced publicly in 2005 by both French president Chirac and German chancellor Schroeder as a public Franco-German initiative to create a competitive European search engine.[67] The Quaero project, which included amongst other members of the European ICT and telecommunications industry the companies Thomson, France Télécom and Exalead, soon lived on as separate German and French public research investment programs, for which state aid was approved by the European Commision in 2007.[68]

2.2.5 Alternatives and the future of Web search services

The consolidation of the Web search industry into an oligopoly or quasi-monopoly of services in the west, does not imply that no alternatives exist or that the dominant services have become the only destination available for Internet users to search for online material. It also does not imply that research and development in relation to Web search only takes place behind the closed doors of a handful of dominant companies. In fact, many small search engine service providers have been developed, many of which still exist. There is a variety of alternatives to the dominant search engines provided by Google or

[64] See also Halavais 2009, pp. 25-26.

[65] See Pandia Search Engine News 2006.

[66] On the French perspective of mobilization against Google, see Jeanneney 2007.

[67] For the speech of the French President, see Chirac 2006. See also Chrisafis 2006.

[68] For the German programme THESEUS, see European Commission 2007b. For the French programme Quaero, see European Commission 2008a.

Miscrosoft to find the location of online material. And if one takes a closer look at these alternatives, one can also often discern competing models for the production of references to online material to Internet users.

First, amongst the alternatives for the dominant search engines there are many so-called 'vertical' or 'niche' search engine providers. These verticals specialize in references to certain types of online destinations. Examples can be found in the context of many specific consumer markets, such as housing, travel or shopping, or with regard to certain types of information, such as medical, legal, financial or geographical data. The ongoing success of many of these vertical search engines is typically attributed to their greater focus in comparison to general, horizontal, search engines, and their resulting ability to select references of high quality for their users on the one hand, and the specific commercial opportunities tied to the matchmaker role between providers of certain goods, services or information and potential users or customers in specific markets on the other hand. Many of these verticals are commercial, but in the public sector we can also find a range of specialized search engines that make specific documents and publicly available information more easily accessible for Internet users. The importance and success of vertical search engines can also be recognized by the various ways in which both Google and Microsoft have acquired, launched and integrated specialized search services into their offering.

Second, there are still alternative horizontal search engines other than Google and Microsoft that have only a limited market share. The most recent example, which emerged after a 3 year long phase of development, is California-based search engine Blekko. Blekko offers end-users a service which is quite similar to the one offered by Google. As mentioned above, in some countries, such as in Russia (Yandex), Czech Republic (Seznam), and further away in South-Korea and China (Baidu), there are strong local competitors.

Third, both academic researchers and entrepreneurs are still actively exploring various alternative models to offer effective means to find online material for Internet end-users. Just one of the interesting alternatives that has been conceptualized over the last decade is a peer to peer model for a Web search engine. There are several academic, commercial, and free and open source software projects that have pursued this model for the production of online references for Internet users. Second, the way in which people use the World Wide Web keeps shifting considerably due to the successful launch of new types of services, such as social network sites (Facebook) and micro-blogging sites (Twitter). These services offer Internet user a different way to select access material in the online environment.

Finally, there are ways in which developments related to online publishing practices more generally could change the search engine environment significantly. An important strand of research and development in the field of Internet information engineering that is strongly related to the thinking about improved search for the Web, is the work on the so-called 'Semantic Web'. The Semantic Web project could be described as an attempt to develop methods and technologies that increase the possibilities for machines to interpret the contents or meaning of online material directly. Consequently, this research focuses more on the improved organization of material on the World Wide Web itself than on improving models to build and operate search engines. Interestingly, the inventor of the Web, Tim

Berners-Lee, is one of the driving forces behind the Semantic Web project, in which researchers and developers are participating since 1999.

From the perspective of Web search engines the Semantic Web project is fascinating for a number of reasons. First, the lack of semantics in the World Wide Web's technical design may have been one of the strongest drivers for the emergence of the Web search industry as we know it. The Web and the hypertext protocols allow any Web publisher to link to anything else. This makes an open online universal document space possible, which is precisely one of the major strengths of the World Wide Web. As we noted earlier in this chapter, the designers of the Web implicitly assumed that Web users would organize the Web. This design philosophy created a strong demand for third party 'useful links' web pages, directories, and search services which would help Internet users to find material located elsewhere.

Second, if we turn to the search engines of today, one could argue that they have in fact developed a kind of Semantic Web overlay, in the sense that search engines have specialized in making recommendations about the relevance, content and meaning of online material based on their own analysis of that material.[69] The big difference is that most of this meta-information about online information is kept behind the closed doors of the server farms that host their version of the annotated index of the Web.

To conclude, the semantic web project, understood as the "*extension of the current [Web], in which information is given well-defined meaning, better enabling computers and people to work in cooperation*"[70] could potentially have a significant impact on the search engine industry if it were to be implemented openly and successfully.[71] While it would allow all search engines to improve their offerings, it could also take some of the power of dominant search engines – the part which is based on their *exclusive* understanding of the material on Web - away from them by opening up similar or even improved meta-data to the Internet community as a whole.

2.3 Conclusion

This chapter has offered a short overview of the history of search engines starting from the early ideas about the opportunities of improved navigation of information in digital information systems to the rise of search engines as one of the most important media of the public networked information environment made possible by the World Wide Web. In particular, it shows how the design of the World Wide Web, which has become the universal platform for online publication since its launch in the early 1990s, implied a natural demand for navigational media and services that would help users find valuable online

[69] See e.g. Arnold 2010.

[70] See Berners-Lee et al 2001. The THESEUS Research and Development investment programme in Germany is specifically focused on the development of semantic Web technologies. See THESEUS.

[71] The Semantic Web project has been quite unsuccessful in making the Web evolve into a semantic web. For a discussion of the utopian project of exhaustive reliable metadata, see e.g. Doctorow, 2001. However, if one understands the semantic Web project as a variety of related methods and technologies that increase the possibilities for machines to interpret the meaning of online material, the Semantic Web project has produced various successes which have been adopted in different contexts. See e.g. W3C, W3C Semantic Web Activity, http://www.w3.org/2001/sw/.

material. Interestingly, although not necessarily surprising, most of these services were initially developed in the academic realm. Later on, the business opportunities related to search engines became an important driver for the further development of the search engine industry and the innovations that have taken place since then, such as the pay-per-click advertisement models that the market leaders use today.

The eventual consolidation of the search engine market can be historically tracked to a number of contributing factors. Some of those factors are inherent in the operation of a general purpose search engine, such as the growing, evolved expectations of search services by end-users or the grown barriers to entry into the market. Other contributing factors include the integration of the search engine as an important business asset in the digital media and ICT industry, thereby reproducing existing consolidation in related markets in the search engine context. Of the many search companies that started offering their service in the 1990s, only Google remains as a mature independent company with its own search engine at its heart.

This points to one of the most remarkable aspects in the history of search engines, namely the fast rise of Google as the dominant global player in the search engine market. In 2011, in many countries, including the Netherlands, Google controls more than 90% of the market in terms of user share and the amount of searches performed on the Web. Not surprisingly, this has attracted a steady stream of commentary over the last decade and has had meant that Google has become synonymous with Web search for the better part of the general audience. But, although Google has had and continues to have an enormous impact on the search engine industry and the way in which Internet users access information on the Web more generally, it is important to look beyond this single company's commercial search service. There are still competitors to Google's search service in the market for general purpose web search, such as Microsoft's Bing. In some jurisdictions strong national alternatives exist, such as in Russia. More importantly, there are numerous other publicly and privately funded services with specific focuses which contribute to the findability of online information for end-users. In addition, research and development in search continues to offer new insights about alternatives to current search engines and the ways in which online information can be organized to enhance effective retrieval of online resources.