Define: Web Search
Semantic Dreams in the Age of the Engine

Master Thesis New Media and Digital Culture | Utrecht University

Student: Shirley Niemans | 0392725

Supervisor: Marianne van den Boomen

Date: November 30, 2009

Abstract

In 2000, Lucas Introna and Helen Nissenbaum argued that search engines raise not just technical, but distinctly ethical and political questions that seem to work against the basic architecture of the Web, and the values that allowed for its growth. Their article was the starting point of a critical Web search debate that is still gaining foothold today. When we consider the semantic metaphor that has been inspiring a refashioning of the Web architecture since 2001, we can see the exact same values of inclusivity, fairness and decentralization reappear that fueled the development of the original WWW. This thesis will explore the 'promise' of the Semantic Web in light of the current debate about the politics of Web search. I will argue that a balanced debate about Semantic Web developments is non-existent and that this is problematic for several reasons. Concluding the thesis, I will consider the dubious position of the W3C in enforcing the implementation of new standards and the power of protocol to be an 'engine of change'.

Acknowledgements

This thesis is a direct result of my research internship at the Institute of Network Cultures (INC), between August 2008 and January 2009. During this internship I was given the opportunity to explore the broad field of Web search using the Institute's elaborate network and the extensive knowledge of the INC staff, and to deliver an editorial outline for the Society of the Query conference. I want to express my warmest gratitude to Geert Lovink, Sabine Niederer and Margreet Riphagen for giving me the incentive and time to start this research, and the trust and support to finish it. You are an inspiration to me.

I'm also indebted to my supervisor Marianne van den Boomen of Utrecht University whose incessant cheerfulness, insightfulness and encouragement have been a fantastic help. Thanks to my friends and brother for (lots of) patience, moral support and pep talks; the word 'thesis' is no longer off-limits. Thesis! Finally and especially, I'm obliged to my parents Leon and Jeannette Niemans whose support on all levels enabled me to start an MA study in the first place.

Table of Contents

Introduction

> "Whenever thought for a time runs along an accepted groove—there is an opportunity for the machine." - *Vannevar Bush, 1945*[1]

The size of the Internet and the amount of information currently available on the World Wide Web (WWW) are of a scale that is hard to imagine. Several studies have attempted measurements but results vary widely. A September 2009 Web Server Survey by Netcraft claims that the Internet holds over 226 million websites.[2] Meanwhile, Google notes to have indexed over a trillion pages with unique URLs – an amount abstract enough to defy all common sense.[3] There is something we do know for a fact however: Search technology has become an absolute necessity for anyone with a need to find meaning and value in the vast amount of information on the Web. This notion finds reflection in the deep integration of the search engine in society today. Google has become our main point of entry to the Internet, and the window through which most Internet users get to see just what the Web is like.

Developments in the field of Web search are rapid. In recent months, news about corporate take-overs, partnerships and the incessant release of new and advanced Web search services have been a daily routine. It seems strange to think back and realize that little over a decade ago, search results did not come with relevance ranking. Scouring through hundreds of search results to find the one relevant to your query, and encountering a host of spam in the process, was not unusual. Until well into the second part of the nineties, human edited directories provided the best alternative to this 'mess'. Although their indices were hardly comprehensive, the links available in these directories came categorized by subject and vouched for by the editors, providing an added value that full-text search engines of the time failed to achieve.

Ranking technologies used by the algorithmic search engines of today have largely fixed the problems of spam and search result overload. The core challenge that guides today's search engine development is known as *disambiguating intent*; figuring out just what the user 'really' means. In the challenge to optimize search engines and offer the best user experience, search result relevancy – the extent to which the first few results match a user query – is key. Google's PageRank algorithm is exemplary in this respect. Unfortunately, the Google window to the Web is not cleanly wiped, the company deliberately sacrifices transparency in order for search results to be free of spam, free of malafide search engine optimization scams and, obviously, in order to sustain its current market share. The engine has in fact become a black box and users are encouraged – and seemingly willing – to trust its deliberations. The near ubiquitous adoption of Google by Web users worldwide brings on new issues in need of addressing. Many scholars have done so inspiringly over the past few years, placing

[1] Vannevar Bush, "As We May Think," The Atlantic (1945).

[2] Netcraft, September 2009 Web Server Survey, 2009 <http://news.netcraft.com/archives/2009/09/23/september_2009_web_server_survey.html> October 23 2009.

[3] Jesse Alpert and Nissan Hajai, "We Knew the Web Was Big..." The Official Google Blog (2008), vol. 2009.

the focal point of their studies on topics such as Google's corporate culture, the politics behind Web search, the influence of search engine use on students' learning abilities, the future of the library, the Googlization of our culture, censorship issues and many more.

An interesting quality of the debate surrounding Web search, is that it invites visions and ideologies from both a Computer Science and a Humanities perspective. Especially Artificial Intelligence (AI) ideals as to what the future of Web search may hold seem fundamentally incompatible with principles held high in a philosophical understanding of what it means to be human. The friction between these views is most apparent in recent discussions surrounding the development of a so-called 'Semantic Web' – the proposed enhancement of the current WWW architecture with a structure of denoted relations between data objects which would make possible 'machine understanding'. Regardless of the (im)possibility of such an endeavor, any change to the Web architecture can reasonably be expected to have an effect on Web search as we know it today; the rapid expansion of the WWW in the last two decades has caused Web search technology to reinvent itself again and again. Interestingly, use cases sketched by the Semantic Web developers in 2001 do not mention search engines, but rather *personal agents* that perform sophisticated tasks combining and integrating data after being instructed by users through their – similarly personalized – Web browsers.

Unfortunately, there is not much literature to be found on the topic of the Semantic Web outside of the realm of the W3C Semantic Web Working Group[4] and computer science in general, where mostly the development of scripting and meta-data languages is discussed. With a 'new' Web seemingly waiting around the corner to change the future of search, this strikes me as odd. As I noted earlier, there certainly are scholars who challenge the very AI-minded image projected by these developments, but a lively and balanced debate similar to the one surrounding the current search engine hegemony, is non-existent. This lack of critical assessment is problematic for several reasons: Firstly, developments are well under way. By the look of things, it is not (and has never been) a question of *whether* the Web ought to be augmented, but *when* the technology will become widely accepted by the Web community. Secondly, the discussion about the idea of the Web as an inclusive space has gained impulse since it is apparent that the dominance of Google affects the visibility and accessibility of online content. With the prospect of an augmented Web, the question of what we want the Web to be like seems urgent once again. Taking these notions into consideration, this thesis will ask:

What is the 'promise' of the Semantic Web, and to what extent do the existing plans for the future of the Web, and the future of Web search in particular, reflect the dominant criticism currently aimed at the search engine?

To answer this question, we will first need to explore the historical context that allowed the Semantic Web to become a 'promise' in the first place. To come to a better understanding of the current state of Web search in relation to the rise of the Web, the *first chapter* will start off with a short history of Web

[4] Ivan Herman, W3c Semantic Web Activity, 2009 <http://www.w3.org/2001/sw/> October 27 2009.

search, its scope ranging from the first pre-Web search engines and Web directories up until the current ubiquity of the algorithmic search engine. I will dig deeper into the indexing, Web crawling and ranking technologies search engines employ, and the core challenges they present today. In the *second chapter*, I will give an overview of the criticism that has developed alongside this historical timeline. As Introna and Nissenbaum argued in their 2000 article 'Shaping the Web: Why the Politics of Search Engines Matters', search engines raise not just technical, but distinctly ethical and political questions that seem to work against the basic architecture of the Web, and the values that allowed for its growth.[5] Their critique is still valid today, and may be extended with the influence of the existing hegemony of one large search engine on traditional information flows. In this chapter I will outline the current 'dominant criticism' and the scope of this debate. The *third chapter* is devoted to gaining a clear understanding of the so-called Semantic Web. What exactly are the proposed technical enhancements, how is this 'semantic', and what are the project's stated goals? Answering the central question, I will address the W3C communication surrounding the developments, as well as humanities scholars mentions of the topic for reflections in respect to the current debate on search engine politics and the future of the Web. Lastly, I will comment on the feasibility of the Semantic Web coming into existence as proposed, in relation to the everyday practice of Web development and the dubious position of the W3C.

[5] Lucas D. Introna and Helen Nissenbaum, "Shaping the Web: Why the Politics of Search Engines Matters," The Information Society: An International Journal 16.3 (2000).

1 Define: Web Search

1.1 Early Web Search

The history of information storage and retrieval is a long one. The origin of the database goes back to the earliest libraries and their qualification systems, and has been accompanied by the need to search and find documents within them. Many techniques for indexing data and defining qualitative criteria for document retrieval have been developed since, some of which failed and some of which are still in some form part of current information retrieval practices. The increased storage power of computers allowed for the advent of the electronic database and of the search engine, a computer program that retrieves links to documents or files from a database, often in answer to a keyword query.

Even before the World Wide Web came into being, search engines existed that sought to organize the information on the many public anonymous FTP (File Transfer Protocol) servers that primarily compiled the network of the early Internet. The first search engine in this context is said to be Archie, created in 1990 by a student at the Montreal-based McGill University. Archie programs regularly downloaded the directory listings of FTP sites, maintaining a searchable file name database.[6] Not long after Archie, Gopher was created at the University of Minnesota, offering a user-friendly menu-based system to pull together distributed Internet resources such as local directories, FTP servers and Wide Area Information Services (WAIS). Various search tools (Veronica, for instance) were built on top of Gopher, allowing for keyword searches of the titles on the Gopher servers. Gopher was eventually overtaken by the technology of the World Wide Web and its hyperlink-based communication protocols that presented a whole new world of interacting with information.[7]

The earliest Web search engines, such as the WWW Worm, retrieved information about Web page URLs, titles and headers using a simple linear search. Search algorithms at that time did not perform link analysis or index the content of a Web page and results were often listed in the order in which they were retrieved, which made it extremely hard to find a document of which one did not know the exact name. With the Web expanding at a fast pace, most of these engines soon slowed to a stop.[8]

With the rise of the World Wide Web also came the first Web directory, installed by WWW creator Tim Berners-Lee: the WWW Virtual Library (VL). The VL is still maintained today as a voluntary cooperation between several topical experts that compile lists of links on their area of expertise. Currently, the VL website states that: “even though it isn't the biggest index of the Web, the VL pages are widely recognized as being amongst the highest-quality guides to particular sections of the Web.”[9]

---

[6] Lee Underwood, A Brief History of Search Engines, 2004 <http://www.webreference.com/authoring/search_history/> October 23 2009.

[7] Cameron Kaiser, "Down the Gopher Hole," TidBits (2007), vol. 2009.

[8] Aaron Wall, History of Search Engines: From 1945 to Google 2007, 2007 <http://www.searchenginehistory.com/> October 23 2009.

[9] The WWW Virtual Library <http://vlib.org/> October 10 2009.

Early Web directories, of which Yahoo! is still the most well known, are modeled largely after the way in which libraries organize collections of books and articles. Editors manually categorize and sub-categorize Web pages by topic. Examples of such categories could be Art & Humanities, Health, Entertainment or News & Media. Sub-categories of Art & Humanities might be photography, literature and history, while Entertainment could be divided in sub-topics such as TV shows, humor, music and games.[10]

For these early directories, comprehensiveness became an issue as the amount of online information grew. The editors often faced considerable backlogs, causing the time between submission of a website and its inclusion in the database at some points to take up to several months.[11] Also, as opposed to the static nature of books, websites tend to be modified regularly, which further frustrated effective categorization and induced 'link rot' as pages were moved or deleted. With the task of maintaining directories becoming harder over the years, the hyperlink structure of the Web turned out to be a short term blessing but a long-term problem: While allowing for links to documents to be listed in multiple categories at the same time, important decisions about the organization of information within the directory system could be avoided.[12] Successfully extending the library metaphor then, seemed less and less probable; rather than how to categorize, the question shifted to whether the categorization method was fitting to the broad and dynamic content of the World Wide Web.

By the mid-nineties, a new approach to Web search was introduced with full-text search. Emerging search engines such as AltaVista and Hotbot started to use computers instead of human editors to download the content of the Web and parse the HTML markup of each page, in order to make an index of all the words (all character groups between two spaces) that a page contained.[13] Making use of a search box on the engine's website, a user would type in a word or a combination of words and the hyperlinks to the documents matching the query were returned. While browsing Web directory categories for interesting content remained an activity many Web users engaged in, Web-savvy users looking for specific documents found a huge improvement in full-text search, and for several years AltaVista was the tool of choice to actually find things on the Web.

For webmasters, the move to full text search significantly changed the development practice. Metadata in HTML markup was first used to enable a company's site to be categorized according to terms such as page title, company name, trademarks, business type and location. Inclusion in the correct category was important for a site to be findable within a directory.[14] Early full-text search engines however, proceeded to parse the metadata as part of the full text to match a user's search terms to the contents of a site. As the data placed inside the HTML <meta> tags would not be displayed on the Web page, webmasters soon found that schemes such as 'keyword stuffing' –

[10] Yahoo!, Yahoo! Directory, Website, Yahoo! Inc. <http://dir.yahoo.com> July 20 2009.

[11] Introna and Nissenbaum, "Shaping the Web: Why the Politics of Search Engines Matters." 171

[12] Win Treese, "Is the Power of Web Search Diminishing?," netWorker 12.2 (2008). 13-15

[13] The Web Search Workshop, Altavista: A Brief History of the Altavista Search Engine, 2009, The Web Marketing Workshop Ltd <http://www.websearchworkshop.co.uk/altavista_history.php> July 20 2009.

[14] Paul Graham, "Metatags -- the Latest Developments," Computer Fraud & Security 2000.10 (2000).

including as many keywords as possible in the meta tag listings – would quite effectively boost their occurrence in search results.[15] In late 1997 and 1998, sometimes referred to as the 'dark ages of search engines', spam became a serious problem for major engines AltaVista and Hotbot.[16] The rapid increase of online information added to this challenge: Too many pages were returned that matched the words from a user query. A vaster quantity of indexed material in this respect attributed little to the search experience without a way of determining which result would be most relevant to the user.[17]

An interesting anecdote exists about computer scientist Jon Kleinberg who discovered a significant weakness in the full-text search system in 1997. When entering the query "search engine" in the popular AltaVista search engine, Kleinberg was surprised to find that the AltaVista homepage was not among its own returned results. Pondering over the reason for the flaw, he soon discovered that the words "search engine" were nowhere to be found on the AltaVista homepage.[18] Following his discovery, Kleinberg set out to develop a system that would be able to effectively rank search results, by giving a value to both the content of Web pages (authority value) and to their links to other pages (hub value). His project was later to become known as the HITS algorithm (Hyperlink-Induced Topic Search), used by a.o. IBM and the moderately successful Teoma search engine.[19]

Kleinberg was not alone in his discovery. Around the same time, Sergey Brin and Larry Page developed the Google search engine and patented their PageRank Algorithm. The key feature of the PageRank algorithm was the way that it treated every link to a Web page as a 'vote', thereby valuing the most linked-to websites highest. Consequently, a link given by a well linked website (a website with high value such as Amazon.com) was worth more than a link given by a less linked-to site such as shirleyniemans.nl. At Stanford University, where Brin and Page started to develop their project, the first Google beta was met with much enthusiasm. The search engine soon gained acknowledgement outside of the University walls and had a profound and immediate impact on Web search. As computer scientist Win Treese remembers, experienced users converted to Google as soon as they had come to try it, increasing its popularity rate very early and very quickly. The engine turned out to be effective without precedent, with the first hit almost always being exactly what you were looking for.[20] The 'I'm feeling lucky' button, a standard feature on the Google website since the 1998 beta, served to illustrate the developers' determination to change the nature of Web search.

In 1998, another directory project went live, too. Supposedly out of discontent with the link rot in Yahoo!'s directory, US computer programmer Rich Skrenta set out to build the largest Web directory. Contrary to Yahoo!, the project that was to become the Open Directory Project (ODP) worked with an unlimited amount of volunteer editors rather than a small team of paid professionals. Better equipped to meet the demands of the expanding Web, it took a year of experimenting and

---

[15] Evan Bleiweiss, "A Brief History of Meta Tags," Media\Outreach (Media Outreach, 2008), vol. 2009.

[16] Brett Tabke, "A Brief History of Seo: The Real Search Engine Wars," WebmasterWorld (2002), vol. 2009.

[17] Treese, "Is the Power of Web Search Diminishing?."

[18] Mike Grehan, "Search, That Was Mighty Sociable," Clickz (Incisive Interactive Marketing LLC, 2008), vol. 2009.

[19] Remus Radu and Raluca Tanase, Lecture #4: Hits Algorithm - Hubs and Authorities on the Internet, 2009, Cornell University <http://www.math.cornell.edu/~mec/Winter2009/RalucaRemus/Lecture4/lecture4.html> July 21 2009.

[20] Treese, "Is the Power of Web Search Diminishing?."

several name changes before Netscape acquired the by then quite extensive directory and announced that the ODP data would be available for anyone to copy and use in other projects.[21] A feature of Web directories is the way links are accompanied by a brief editorial note, describing the content of the page and often its relevance within the (sub) category. Today, projects such as the Virtual Library have long given up the idea of being comprehensive and focus instead on the expertise of the editors and the quality of the compiled links. The ODP targets both quantity and quality as it now claims to be 'the largest, most comprehensive human-edited directory of the Web", allowing net-citizens to "each organize a small portion of the web and present it back to the rest of the population, culling out the bad and useless and keeping only the best content."[22]

The mutual relationship between contemporary Web directories and search engines is somewhat of a mystery. Clear information about the subject is difficult to find, especially on the part of the search engine. The ODP directory announces that it: "powers the core directory services for the Web's largest and most popular search engines and portals, including Netscape Search, AOL Search, Google, Lycos, HotBot, DirectHit, and hundreds of others."[23] The search engines mentioned however – the ones that still exist, that is – no longer list their directories as a service on the search engine homepage. The Amsterdam-based Digital Methods Initiative (DMI) conducted interesting research on this topic, which resulted in a video of the Google interface between 1998 until late 2007.[24] The video shows the waning status of the Google directory link as it moved from being a visible front-page tab between 2000 and 2004 to a secondary position in the 'more' section, after which it was moved even a further click away to the 'even more' section. For a period of time even, the Google directory was accessible only by performing a Google search. According to the researchers, the sad fate of the directory link illustrates the demise of human librarian work in organizing Web content, in favor of the back-end algorithm.

Considering, then, this rough timeline of Web search we can see that the categorizing library model as such seems to have lost from the algorithmic, popularity-based ranking methods in providing the most efficient search experience on the ever-expanding World Wide Web. Editorial opinion has not been eradicated from the engine completely, although the extent to which this affects search results is for the best part unclear. Today, Google is the most popular engine occupying a staggering 90% of the global Web search market, followed only at great distance by Yahoo! with a 4.3% market share, Bing occupying 3.2% and the Chinese search engine Baidu that takes up 0.47%.[25] Search engine technology has advanced significantly over the last decade, while at the same time surprisingly little is known, or even knowable, about the exact way in which the engine processes a query, and how the

[21] Search Engine Laisha's List, Feature: Odp History, 1999 <http://www.laisha.com/zine/odphistory.html> October 23 2009.

[22] DMOZ, About the Open Directory Project, 2009 <http://www.dmoz.org/about.html> October 23 2009.

[23] DMOZ, About the Open Directory Project.

[24] Digital Methods Initiative, The Demise of the Directory: Web Librarian Work Removed in Google, 2008 <http://wiki.digitalmethods.net/Dmi/DemiseDirectory> October 23 2009.

[25] StatCounter Global Stats, Top 5 Search Engines from Sep 08 to Oct 09, 2009 <http://gs.statcounter.com/#search_engine-ww-monthly-200809-200910> October 23 2009.

order of result listings has come to be defined. In the next paragraphs I will dig a little deeper into the crawling, indexing and ranking processes of the largest search engines, focusing firstly on the common characteristics and secondly on the qualities by which the engines compete.

1.2 Search Engine Technology

The mechanics behind most general-purpose search engines share a common structure. When a user enters a set of keywords in a search box on a website, the Web server queries the search engine index which contains information about Web pages from all over the Web, such as which word occurs where, as well as some system of determining the relative value of each page. After processing this information, the query moves to the document database where copies of the original pages are stored. Snippets (a few lines of text containing the queried keywords) of matching pages are retrieved and presented to the user in a ranked list of search results.

The engine's document database is compiled and maintained by Web crawlers, also known as spiders or software (ro)bots, which continuously register, call up, and archive Web pages by following a repeated series of steps. Contrary to what the name suggests, crawlers don't actually travel the Web, but they maintain a long list of URLs that either are indexed already or are known to exist. From this list, the crawler selects a URL that has not been indexed and retrieves the page for analysis (parsing) by the indexing program. An index record is then created for the page and added to the overall index. The indexing program also extracts all hyperlinks to other Web pages found in the document and adds these to the list of URLs for the crawler to retrieve in the future.[26]

Crawlers exist in many types, and are modified regularly in order to better reach the 'deep Web', the part of the Web that is notoriously difficult to index by search engines. However advanced the technology, there are still types of material that search engine crawlers are unable to index. A large part of the so-called deep Web or 'invisible Web' is made up of the content of searchable databases that either require a manual login, or present results only in response to a user query.[27] For website owners, it is possible to control crawler behavior to some extent by using a robots.txt file. Before crawling a site, most automated bots are programmed to search for a robots.txt file in order to check whether the document contains guidelines as to which pages may and may not be accessed.

Figure 1 visualizes what Google terms the 'Life of a Google Query'. Concerning the index servers, Google states that their contents closely resemble the index one would find at the end of a book. There are significant differences however that make this a somewhat misleading comparison. In order to compile its index, Google takes the 'raw' crawling data (a huge amount of sites) and creates a long list of all the different words that occur on those sites. Along with each word, the documents that contain that specific word are registered. When a user runs a query for 'civil war' for instance, two posting lists are compiled that list the documents containing the word 'civil' and the documents that contain the word 'war'. From these lists an 'intersecting posting list' is deduced in which only the

[26] William Y. Arms, Digital Libraries (Cambridge: MIT Press, 2001). 710-711

[27] UC Berkeley, Invisible or Deep Web: What It Is, How to Find It, and Its Inherent Ambiguity, 2009 <http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/InvisibleWeb.html> October 23 2009.

documents sporting both words are left.[28] While this is just the first step in defining the results a user gets to see, the indexing process itself is still based on full-text search, harvesting all words in a document rather than only the ones relevant to the subject a document deals with – as would the index in the back of a book. The fact that a search for civil war does in fact render useful results, and not just documents on any war with an added mention of the word civil, is due to the elaborate relevance metrics that Google employs, which I will discuss in more detail later.

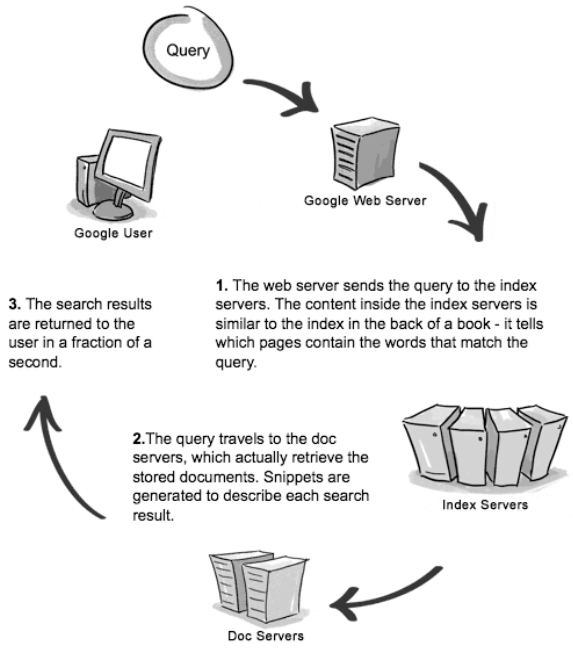

Figure 1. Life of a Google Query[29]

Contrary to the index-based search engines that emerged in the mid-nineties, (re)crawling and indexing today takes place continuously. Web content has become increasingly dynamic in the past few years with an incessant stream of user generated content being uploaded to the Web, and blog posts losing momentum when indexed too late. Search engines experience hardly any backlog lately, as an experiment by Google's Matt Cutts illustrates: Proving his point by taking screenshots that included his desktop clock, he shows that Google indexed his latest blog post within half an hour after he hit the publish button.[30]

Considering the continuous expansion of the Web, it is hard to measure how many documents – and unique ones at that – are actually there to be indexed. Regardless though, search engines have been quick to baffle both users and competing engines by advertising the huge number of documents they

[28] Matt Cutts, How Does Google Collect and Rank Results?, 2009, Google <http://www.google.com/librariancenter/articles/0512_01.html> July 21 2009.

[29] Google, Corporate Information: Technology Overview, 2009, Google <http://www.google.com/corporate/tech.html> July 21 2009.

[30] Matt Cutts, "Minty Fresh Indexing," Matt Cutts: Gadgets, Google, and SEO (2007), vol. 2009.

have been able to retrieve. In September 2005, on the verge of its seven-year anniversary, Google stopped boldly announcing the amount of Web pages it indexed on its homepage – which it claimed to be a total of 8,168,684,336 at the time.[31] This decision followed Yahoo!'s claim a month earlier that it by then indexed 19.2 billion documents on the Web, a number vast enough to cause outrage among Google's executives and to inspire a renewed debate about the accuracy and meaningfulness of such numbers.[32] First of all, it is unclear by which metrics both Google and Yahoo arrived at their assertions. One may ask for instance who decides what content is taken to count as a 'page', as the metaphor is starting to feel outdated on the contemporary dynamic Web. Dynamic websites such as Google Maps might in theory generate an infinite amount of pages, and the millions of sound, video and image files on the Web have unique URLs as well. It has been pointed out by John Battelle that there exist no benchmarks for these measurements that may allow a less subjective third party to once and for all decide the debate.[33] Furthermore, as developments within full-text search have illustrated, the size of the index is but a small part of what constitutes a useful search engine.

Figure 2. The Google Homepage on September 25, 2005 (left)[34]

Figure 3. The Google Homepage on September 26, 2005 (right)[35]

1.3 Ranking Schemes

The battle between the engines continues over the quality of search results although, as we will see, large figures continue to play a role in the way search engines describe their process. The relevance of search results and the weight attributed to individual Web pages have proven to be important tools for engines to distinguish themselves. Google demonstrates such quite successfully with its patented

[31] The Internet Archive, http://www.google.com September 25, 2005, Internet Archive Wayback Machine <http://web.archive.org/web/20050924172505/www.google.com> July 21 2009.

[32] John Battelle, "Google Announces New Index Size, Shifts Focus from Counting," John Battelle's Searchblog (2005), vol. 2009.

[33] John Battelle, "In This Battle, Size Does Matter: Google Responds to Yahoo Index Claims," John Battelle's Searchblog (2005), vol. 2009.

[34] Internet Archive, http://www.google.com September 25.

[35] The Internet Archive, http://www.google.com September 26, 2005, Internet Archive Wayback Machine <http://web.archive.org/web/20050926235645/www.google.com> July 22 2009.

PageRank algorithm. On the subject, the company's Corporate Technology Overview states the following:

> "PageRank reflects our view of the importance of web pages by considering more than 500 million variables and 2 billion terms. Pages that we believe are important pages receive a higher PageRank and are more likely to appear at the top of the search results. PageRank also considers the importance of each page that casts a vote, as votes from some pages are considered to have greater value, thus giving the linked page greater value. We have always taken a pragmatic approach to help improve search quality and create useful products, and our technology uses the collective intelligence of the web to determine a page's importance."[36]

Google thus attributes higher PageRank to the sites it believes to be more important, based on an enormous amount of algorithmically applied variables. The emphasis on quantity here is interesting, especially since the number emphasizes the giant scope of things the average user does not get to know. Of course some of these variables are no secret; the company provides elaborate guidelines and tools for developing 'search engine friendly' websites on its Webmaster Support page. A certain amount of transparency in this respect is needed to inform developers of where to put important information that will allow the crawler to analyze pages effectively. The amount of variables mentioned however is large enough to leave no doubt about the overall unknowable nature of Google's ranking process, posing a major challenge to the Search Engine Optimization (SEO) industry.

The practice of getting sites to reach top five ranking in Google – in either a search engine friendly or unfriendly manner - has become a lucrative business over time. In many types of trade, a company's success is directly related to its visibility on the first page of search results. The SEO industry encounters some hurdles however; Google is extremely keen on optimization and punishes sites it suspects to have achieved their ranking in dubious ways. On the optimization topic, Google's Webmaster Tools directory states: "While SEOs can provide clients with valuable services, some unethical SEOs have given the industry a black eye through their overly aggressive marketing efforts and their attempts to manipulate search engine results in unfair ways. Practices that violate our guidelines may result in a negative adjustment of your site's presence in Google, or even the removal of your site from our index."[37]

As for the user, Google's stated pragmatism is not without benefit. The engine has successfully managed to eradicate spam from its search results and developers are likely to think twice before spiking their PageRank in an unwarranted manner. While emphasizing the non-transparent nature of its ranking mechanics seems hardly a thing to boast about – in the case of Google the user is encouraged to 'blindly' trust the search giant in its considerations.

---

[36] Google, Corporate Information: Technology Overview.

[37] Google Webmaster Central, Search Engine Optimization (Seo), 2009 <http://www.google.com/support/webmasters/bin/answer.py?answer=35291&cbid=1qcfyilhw8zwr&src=cb&lev=answer> October 23 2009.

Building forth on the Google formula, runner up search engines by Yahoo! and Microsoft have updated their strategy toward Web search over the last few years, quite openly countering several of the assumptions underlying Google's PageRank technology. In January 2008, Yahoo! patented its 'User Sensitive Pagerank', a technology that adds user behavior as a variable to the ranking process.[38] The User Sensitive Pagerank algorithm gathers large amounts of data about the actual use of websites. This information, such as the duration of visits or the number of times a link gets clicked, is then put to use in determining the weight that is attributed to links and the value that is given to websites. Yahoo! claims that by reflecting "the navigational behavior of the user population with regard to documents, pages, sites, and domains visited, and links selected"[39] in search results, User Sensitive Pagerank takes the differences between links into account rather than their technical similarities. In this way, the technology is said to be able to distinguish between old, new, popular and hardly used content, determining a so-called 'authority value' for Web documents.

Microsoft has taken another approach recently by launching Bing[40], the Live Search follow-up, on June 3rd 2009. Marketed as a 'decision engine', Bing focuses not so much on the analysis of user behavior, but on an alleged improvement of user experience. Microsoft states: "Bing is a new search engine designed to do more than merely help you find information. Bing organizes search results and provides refinement tools that help you overcome information overload, get things done and quickly bring you to the point of using that information to make an informed decision."[41] Microsoft seems to target both Google's minimalist and its universal mantra, adding an extensive array of options and tools that organize data on the interface level, as well as focusing on a smaller amount of 'vertical', or specialist search markets: product comparisons, travel planning, health research and finding local businesses.

The search engine landscape gained impulse recently when Yahoo! and Microsoft announced a 10-year partnership in the fields of Web search and advertising. After more than a year of deliberations, take-over attempts and ultimately a Yahoo! management shift, the two companies have joined forces to better compete with the current Google hegemony.[42] In the mean time, Google has announced a forthcoming update of its search engine, nicknamed 'Google Caffeine', which would in turn render Bing obsolete.[43] Preliminary try-outs of the Caffeine sandbox have not rendered too many changes in search result rankings, although some findings noted the new Google to have extended its

[38] Pavel Berkhin, "User-Sensitive Pagerank," ed. US Patent & Trademark Office (YAHOO! INC., 2008), vol. 474195/11.

[39] Berkhin et al., "User-Sensitive Pagerank."

[40] Microsoft, Bing, 2009, Microsoft <http://www.bing.com> July 21 2009.

[41] Microsoft, About Bing, 2009, Microsoft Corporation <http://www.discoverbing.com/behindbing/about.aspx> July 21 2009.

[42] The New York Times, Yahoo-Microsoft Deal, July 30 2009 <http://topics.nytimes.com/top/news/business/companies/yahoo_inc/yahoo-microsoft-deal/index.html> October 23 2009.

[43] BBC NEWS, New Google 'Puts Bing in Shade', August 12 2009 <http://news.bbc.co.uk/2/hi/technology/8195739.stm> October 23 2009.

interface options, privilege fresh content and to show an inclination toward 'real-time search'.[44] Tapping into the real-time Web, the part of the Web compiled by status update and micro blogging services like FriendFeed and Twitter, has proven valuable as it connects search to what is happening right now, anywhere in the world, on any given subject. At this point within recent developments, it is yet unclear whether the Microsoft and Yahoo! partnership, as well as Google's update will notably change the Web search experience.

1.4 Ambiguous Intent

The core challenge that contemporary search engines face is known as 'disambiguating intent'; finding ways to diminish the extent of ambiguity within a user query. As noted earlier, queries are highly contingent. Circumstances, experience and further context that defines user intent is lost in the process of translating and abstracting human questions into a keyword query. It is near impossible to determine what a user really means when she enters a keyword query into a search box. Vice versa, it is equally difficult for a user – even if she would know exactly what she wants – to clearly communicate intent toward the engine, especially when faced with limited knowledge about the ways of the algorithm.[45] Due to the current workings of ranking algorithms employed by the larger search engines, the heuristics used for determining what a user means by a certain query are biased by popularity, which in the case of a search for "Paris" might result in several encounters with the celebrity before finding the intended pages about the capital of France.

The success of vertical search engines amounts to the fact that it is easier to disambiguate user intent when the query concerns a constrained topic. Vertical search engines focus not on the horizontal breadth of information and topics on the Web, but slice through it vertically as it were, zooming in on one topic and offering a more targeted search experience. For a search engine that specializes in online shopping for instance, it would be fairly safe to assert that a user entering a query for 'iPod' is not looking for hardware support on one she already owns, but is looking to buy, compare prices or find stores.

On other engines, such as newcomer Wolfram|Alpha, only very specific types of searches can be performed. Wolfram|Alpha, marketed as a 'Computational Knowledge Engine' focuses on building a database of systematic knowledge that can be computed to generate output: "Our goal is to build on the achievements of science and other systematizations of knowledge to provide a single source that can be relied on by everyone for definitive answers to factual queries."[46] The engine fully relies on its own internal knowledge base, and does not search the Web or return links. Computational knowledge in this respect refers to quantitative and widely accepted facts, such as mathematical models and statistics on the level of demographics, geography, physics and economy. Queries fitting to the computational engine would for instance be "How many protons are in an oxygen atom?", "What is

[44] BBC NEWS, New Google 'Puts Bing in Shade'.

[45] Larry Cornett, "Search & Serendipity: Finding More When You Know Less," Search Engine Land (2007), vol. 2009.

[46] Wolfram|Alpha, About Wolfram|Alpha, 2009 <http://www.wolframalpha.com/about.html> October 23 2009.

the average depth of the North Sea?" or "x^2 sin(x)". Consequently, Wolfram|Alpha does not work for non-computable, subjective or 'fuzzy' data where there is no one correct or widely accepted answer to be computed by the system. Using the constrained collection of its database only, Wolfram|Alpha aims to function as a primary source, as opposed to secondary sources that offer links to original content, such as Google or Wikipedia.[47]

The challenge toward minimizing ambiguity fuels many of the developments in the Web search industry. With Bing, Microsoft's strategy seems to be a rich interface, offering many ways in which a user may refine her initial search or find related searches. Google's answer to ambiguity of intent has initially been universal search; blending results from its video, news, images and book search engines among its Web crawling results. Today however, signing up for a Google account and searching as a 'logged in user' increases the possibilities to refine the presentation of one's personal search results. With the 'Web History' feature, Google makes use of a user's earlier searches and previously selected search results to predict which pages are likely to be of more personal relevance. Furthermore, signed in to Google it is possible for a user to promote a result that she believes should be rated higher according to the specific query, to remove irrelevant results and to add comments to URLs.[48] Currently, these features apply to the specific user account and search results only, although it is not hard to imagine Google eventually using these personal votes in some way to improve on its overall ranking process.

Less specifically targeted is the practice of using the IP address of the computer a search is performed on to make assumptions as to the background of the user. Even when a user is not logged in to a search engine or when privacy concerns prohibit the use of personal data, IP addresses can be parsed to provide variables such as the geographical location of the computer and the local timeframe of a search.[49] There is evidence of Google applying such personalization techniques on non-identified users: On his Web log, Siva Vaidhyanathan has noted that a search on the Indian version of Google (Google.in) performed in Charlottesville, Virginia, will render a different result set than when the same search is performed on the Indian Google from a computer located in New Delhi.[50]

The techniques used to disambiguate user intent and enhance the search experience are also cause for worries, as they may as well be put to use for alternative purposes as search engine services expand. The parsing of IP addresses and the storage of user information and search results for an extended period of time raises concerns about the violation of users' (digital) civil rights. As is currently the case in China, restrictive governments may instruct search engines to block certain content requested by computers within national perimeters. Personal search histories may be used in a similar

[47] Wolfram|Alpha, Frequently Asked Questions, 2009 <http://www.wolframalpha.com/faqs.html> October 23 2009.

[48] Google Web Search, Features: Searchwiki, 2009 <http://www.google.com/support/websearch/bin/answer.py?hl=en&answer=115764> October 23 2009.

[49] Qiaozhu Mei and Kenneth Church, "Entropy of Search Logs: How Hard Is Search? With Personalization? With Backoff?," Proceedings of the international conference on Web search and web data mining (Palo Alto, California, USA: ACM, 2008.

[50] Siva Vaidhyanathan, "Another Chapter: The Many Voices of Google," The Googlization of Everything (2009), vol. 2009.

way to detect and locate users that violate national laws and regulations. On another note, the ranking mechanisms employed by search engines and the emphasis on popularity as the prime incentive have caused links on the Web to be distributed in a very specific way, affecting the overall visibility and accessibility of online content. Considering all of the above, questions arise as to whether the values with which the Web was originally created still find reflection on the Web today. The next chapter will take a deeper look at the way in which the debate about the politics of search engines is taking shape.

2 Search Politics

In the much cited article “Shaping the Web: Why the Politics of Search Engines Matters”, written in 2000, Lucas Introna and Helen Nissenbaum argue that the advent of the search engine raises not only technical questions, but distinctly political ones. At the time of their writing, the Web had come to show its potential and many predictions were made as to where the new medium was headed, particularly in relation to traditional media. Would it evolve to be a largely commercial network most beneficiary to corporate ambition, as indicated by US vice president Gore at the time? Or was the Internet essentially – as Mark Poster put it – a postmodern and democratizing medium that defied authority and encouraged dissemination rather than centrality? In their article, Introna and Nissenbaum emphasize the Web's potential to be of public good, but state that several political, economical and technical factors condition this potential.

In 2000, two categories of search engines could be distinguished; the ones that used spiders or bots to index the pages, such as Alta Vista, Lycos, or Hotbot and the ones that relied on human editors for indexing Web pages (directory-based search engines such as Yahoo! at the time). As indicated in the prior chapter, in both cases it was important for an estimation of the value of search results to make some judgment in the indexing process, whether this is done by human editors or algorithms. However, the core concern put forward by the article is: *What guides the human and what guides the spider*? Introna and Nissenbaum specifically note the tendency of larger search engines to give prominence to popular content and ‘wealthy’ sites at the expense of others, thereby undermining the substantive ideal of the Web as an inclusive space.[51]

2.1 The Web as an Inclusive Space

There are different levels of abstraction on which one may observe the Internet and its architecture, as studies in network topology have shown. Many maps can be drawn showing the nodes and relations in a network, inviting various assumptions about the nature of the medium depending on where the lens is aimed. For example, one may sketch the physical architecture of the Internet on the level of the fiber-optic cables that connect computers, domain name servers and routers, or one may sketch the level of the World Wide Web; the documents and multimedia content connected by hyperlinks and the data flow facilitated by the communication protocols. When the World Wide Web was in its first stages of development in the early 1990’s, Tim Berners-Lee described it as:

[51] Introna and Nissenbaum, "Shaping the Web: Why the Politics of Search Engines Matters." 170-185

> "A seamless world in which ALL information, from any source, can be accessed in a consistent and simple way. […] To allow the web to scale, it was designed without any centralized facility. Anyone can publish information, and anyone (authorized) can read it. There is no central control. To publish data you run a server, and to read data you run a client. All the clients and all the servers are connected to each other by the Internet. The W3 protocols and other standard protocols allow all clients to communicate with all servers. […] Once information is available, it should be accessible from any type of computer, in any country, and an (authorized) person should only have to use one simple program to access it."[52]

On an ideological level, the idea of the Web as an inclusive or democratic space is very strong and seems to underlie debates about the future of the Internet even today. Berners-Lee's description emphasizes the fact that on the level of the architecture, the Web is inclusive and democratic by nature: There is access for all, anyone can read, set up a server and publish, there is no centralized facility and no central control as the communication protocols distribute data in small packets across the network. While these features are indeed hard to contest it may be argued, especially since the advent of the search engine, that inclusiveness in this respect says little about the visibility and accessibility of content on the Web.

In 1999, Albert-Lázló Barabási and Réka Albert discovered that in the complex network of the World Wide Web, the connectivity between nodes (websites) follows a power-law distribution, also known as the 80/20 rule. Prior to that time, the network was assumed to be random - that is to say, it was assumed that links were distributed randomly across the network. In reality however, 20% of all the nodes has a huge amount of links - forming important hubs – while the vast majority of nodes has only a few links. This manner of distribution is a consequence of the fact that people do not link randomly. In fact, new websites are far more likely to link to popular sites – the ones that are already wealthy in number of links. In a continuously growing network such as the Web, this is cause for the 'rich get richer' phenomenon, in which popular sites gain popularity as the network expands.[53] Also, as opposed to networks where ties between nodes go both ways – such as the fiber-optic cables that allow traffic to and from each computer they connect – hyperlinks are one-directional: hyperlink points toward one URL and there is no guarantee that, once arrived on this webpage, there exists a link back to the original site. This makes the Web a so-called directed network. Rather than forming a single homogeneous whole, directed networks break up into different parts with there own set of 'traffic rules'.[54] As a consequence of the directedness of links, mapping out or indexing the entire Web is an impossible task that may only render an incomplete picture.

Search engines tackle part of this indexing difficulty by asking webmasters to submit sites to their indices manually, which allows the robot to index parts of the Web that it would normally not have

[52] Tim Berners-Lee, Worldwide Web Seminar, 1991 <http://www.w3.org/Talks/General.html> July 27 2009.

[53] Albert-László Barabási and Réka Albert, "Emergence of Scaling in Random Networks," Science 286.5439 (1999).

[54] Albert-Laszlo Barabasi, Linked: How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life (New York: Plume Books, 2003). 168

access to.[55] In the process of expanding the index, crawlers – or spiders in Introna and Nissenbaums terminology – and humans work in cooperation. As noted earlier however, the issue of the visibility and accessibility of online content does not end with the spider or with the search engine's index size. In the current Web search ecosystem, it is especially through the employment of ranking mechanisms that search engines do not transcend, but instead significantly add to the centralizing notion on the Web.

2.2 The Centralizing Web

Early studies of information politics on the Web, such as those presented in the edited volume *Preferred Placement* (2000), challenge the idea of the Web as a level playing field, cooling some of the initial euphoria in the discourse surrounding the democratic and public sphere potential of the medium. Instead of focusing merely on the possibilities of the Web - the dominant approach at the time - a need for research into the constraints of the medium is expressed. Specifically, the extent to which users can be said to individually or collectively 'author stories' by their online behavior is questioned and attention is redirected to the technological features that either limit or promote the paths that are there to be taken.[56]

The decentralizing or centrifugal force of the Web, existing in the fact that anyone may publish, be it 'amateur' or 'expert', inspired the idea of a level playing field over which established institutions would hold no power. Mark Poster's vision of a postmodern and democratizing medium that defied authority and encouraged dissemination rather than centrality is exemplary in this respect. Countering the centrifugal force however, is the centralizing tendency to the distribution of links on the Web that finds reinforcement in search engine workings. As we have seen in the former chapter, contemporary search engines organize a large part of their search result ranking around the practice of linking. Google takes the idea of links as popularity votes quite literally; sites that are most linked-to are considered more important than sites that are less well linked-to. As a consequence, the well-linked sites rank higher on the result pages and have proven to receive more traffic from search engine users. The fact that users tend to refrain from looking at more search results than the first few pages on offer, commonly referred to as 'link laziness', only strengthens this effect.[57]

Nicholas Carr, a well-known critic of technological utopianism, remembers how the Web seemed new and liberating in the 1990s, promising to level the media playing field once and for all. Large media players were slow to make their move to the Web, but they did, and their gravitational pull proved strong as ever:

> "Even back then, the counterforce to the web's centrifugal force - the centripetal force that would draw us back toward big, central information stores - was building. Hyperlinks were

[55] Barabasi, Linked: How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life. 166-169

[56] Richard Rogers, ed., Preferred Placement - Knowledge Politics on the Web (Amsterdam: De Balie, 2000).

[57] Nicholas Carr, "The Centripetal Web," Rough Type (2008), vol. 2009.

> creating feedback loops that served to amplify the popularity of popular sites, feedback loops that would become massively more powerful when modern search engines, like Google, began to rank pages on the basis of links and traffic and other measures of popularity. Navigational tools that used to emphasize ephemera began to filter it out. Roads out began to curve back in." [58]

Attention on the Web is distributed differently as opposed to earlier information environments. Sure enough, as Chris Anderson's 'long tail' theory illustrates, fringe content is there in vast quantities, but hyperlink analysis has shown that the Web isn't too supportive of the diversity of the content it holds.[59] Through ranking technology, search engines create winners and losers, reinforcing traditional authorities in the process.

At the same time, this line of thought suggests that alternative ranking processes may encourage a move away from mainstream content. In June 2008, Wikipedia co-founder Jimmy Wales launched the open source search project Wikia Search, an attempt to re-democratize and open up the search engine process by allowing the user community to rearrange search result listings, evaluate and comment on individual links, add links and remove those that did not match the query.[60] Several months after the official launch (a beta had been running since January 2008), Wikia was reported to average about 50,000 queries a day from over 391,000 registered users. These were numbers that hardly impressed the larger search engines, but the community formula, a large part of Wikipedia's success, did seem like a force to reckon with.[61] On March 31 2009 however, Wales announced that due to the merely moderate success, the company had decided to concentrate its resources on more successful Wikia projects such as Wikia Answers.[62] A very recent start-up is peer-to-peer real-time search engine Wowd, launched on October 20, 2009. Its approach differs from most engines in that it relies on users to download and install a peer-to-peer client that records users' real-time click stream data (if one allows it to) to both index and rank search results. The index then isn't located on a central server cluster, but is scattered across the users' machines and results are ranked according to how often and how recently a site has been visited.[63] Up and running for just a couple of weeks now, there are no useful results to take into account yet.

Compared to the much stronger pull of the centre however, magnified by billions of user transactions a day, the disseminating force on the Web is still the weaker quality. A 2007 study by Huang et al. has established that more experienced or heavy Web users tend to concentrate their

[58] Carr, "The Centripetal Web."

[59] Chris Anderson, "The Long Tail," Wired 2004.

[60] Torsten Kleinz, "Wikia Search Opens Its Doors," The H (2008), vol. 2009.

[61] Clint Boulton, "Google Search Ranking Feature Threatens Wikia Search," eWeek (2008), vol. 2009.

[62] Jimmy Wales, "Update on Wikia – Doing More of What's Working," Jimmy Wales (2009), vol. 2009.

[63] Erick Schonfeld, "Wowd Takes a Stab at Realtime Search with a Peer-to-Peer Approach," TechCrunch (2009), vol. 2009.

online information behavior on a core set of websites that function as anchors.[64] While heavy or experienced Web users may be thought to be less affected by a search engine's bias toward a global consumer culture, the existing hegemony of the Google search engine does pose a threat to cultural diversity as more cultural content moves online. Or, as Nicholas Carr would have it: "The long tail is still there, of course, but far from wagging the web-dog, it's taken on the look of a vestigial organ."[65] The matter of cultural diversity also extends to the way in which search engines manage to reflect the broad range of international material available on the Web. The US domination of the search market and the tendency toward English resources on the Web have been cause for Jean-Noël Jeanneney, president of the French National Library at the time, to make a pressing case for search engines that are created and managed in Europe.[66] The 'Googlization' of cultural artifacts furthermore reduces the ability of governments to protect less popular resources, he argues, and as such to protect diversity. PageRank-wise, more links lead to US pages; firstly since they have often existed longer, and secondly since early winners keep their advantage - a mechanism that is again multiplied by search engine workings.

2.3 The Google Hegemony

The critique Introna and Nissenbaum envision in their original article is valid even today, and may be extended with the influence of the existing hegemony of one large search engine on traditional information flows and the distribution of power. Search engines raise political concerns not only because of their technical workings, but also since the manner in which they do so does not accord with the idea of the Web as a public good; both in the sense of being a collaborative space as well as in its incarnation as a conveyor of information. There is not much use to the Web and to the vast amount of information it holds when there are no useful ways to actually search its contents. In the current attention economy, 'inclusiveness' translates to visibility and accessibility through search means, and it is in this respect that the engine as the dominant search paradigm holds a significant amount of political power.

An important aspect of the PageRank technology is the distributed responsibility of ranking. In the transition from full-text search to PageRank, hacking a site's position within the search results by using keyword stuffing was rendered useless, since link popularity became the principle ranking mode. Although this mode was discovered to be hackable at first too, Google managed to eradicate malafide SEO schemes fairly quickly. The change toward the algorithmic engine was a cultural one, too. Google's advantage at a time when the Web was growing exponentially soon made it *the* Web search standard. In a manner of speaking, Google's politics introduced the law to the 'Wild West' of Web search; non-compliance to its standards was met with exclusion from the index - the equivalent

[64] Chun-Yao Huang, Yung-Cheng Shen, I-Ping Chiang and Chen-Shun Lin, "Concentration of Web Users' Online Information Behaviour," Information Research 12.4 (2007).

[65] Carr, "The Centripetal Web."

[66] Jean-Noël Jeanneney, Google and the Myth of Universal Knowledge (Chicago: University of Chicago Press, 2007).

of being rendered completely invisible.

Another aspect of the power of Google is introduced as 'The Google Dilemma' in a recent article by James Grimmelmann in the New York Law School Review. Grimmelmann introduces the subject by taking as a case study two well-known cases of 'Google bombing': 'Talentless Hack' and 'Jew'. Google uses link analysis not only to determine the importance of an individual page but also to determine what the page is about, by looking at the phrases people use to describe a link. It so occurred that in 2001, college student Adam Mathes asked a bunch of friends to link to another friend, Ben Brown, and to add the description 'talentless hack'. Soon after, Ben Brown would be the first hit in a search for 'talentless hack' on Google. A much more serious case of Google bombing occurred around the search term 'Jew'. In 2004, the first hit for the keyword would be the anti-Semitic jewwatch.com website, featuring an abundance of hateful material. A Jewish activist then set off a successful Google bomb by asking people worldwide to link the word Jew to a Wikipedia article. The response was vast enough for a neo-Nazi counter-Google bomb to sort no effect.[67]

Along with the Wikipedia Google bomb, however, came a petition to have Google remove jewwatch.com from its search results entirely - a request to which Google did not comply. Obviously, Google possesses the technical means to remove the search result once and for all, but instead the company decided to post a notice to the Jew search result page, explaining how its employees were also disturbed about the offensive nature of these search results. In his article, Grimmelmann notes how 'don't intervene in search results' seems even more of a motto to Google than 'don't be evil', seeing that the search giant "just feeds the Internet into its server farm and waits for results to emerge, unsullied by mere mortals. Thus, the "Offensive Search Results" page that Google put up to explain why 'Jew' takes you to Jew Watch is Google's' way of saying "don't blame us, the computers did it"."[68]

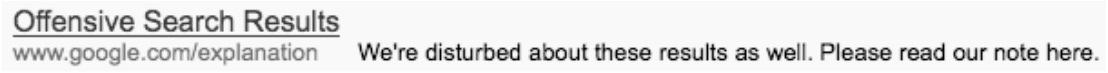

Figure 4. Google notice in the sponsored links section with search results for "Jew"

The question then arises to what extent Google, or any search engine, is automated as to reflect no opinion of its employees or programmers in its search results. More specifically put perhaps; why would the company hold this idea so highly? Although the metrics are largely unknown, the development of Google's algorithms has doubtlessly involved numerous decisions about ranking criteria by staff and developers. In order to deliver the most relevant results and to eradicate spam and malafide ranking schemes, PageRank gets updated all the time in response to new developments and evolving user needs. James Grimmelmann refers to the Google Dilemma as the tension between this emphasized 'objective' nature of computed results and the beliefs and preferences of its staff, which he considers proof of Google's uncomfortability within its position of power. As far as the broad content of the Web goes - as opposed to confined data collections - there is a strong argument to be

[67] James Grimmelmann, "The Google Dilemma," New York Law School Review 53.939 (2009).

[68] Grimmelmann, "The Google Dilemma." 944

made for the claim that much like Wolfram|Alpha, Google does not work well for 'fuzzy' data either. Questions as to what it means to be a Jew or what search results for 'Tiananmen' should reflect in China are difficult political and cultural issues, and responsibility is carefully redirected toward the machine or toward 'overriding' national laws.

Google's many ways of modeling the access to information are a constant source for criticism. In his 2009 book *Search Engine Society*, Alexander Halavais defines the core search engine question in this respect as: "Who sees what, under which circumstances and in what context?".[69] From an early stage on, national borders have been enforced online, and search engines serve as an important site of control and surveillance across the globe. It is the area of search engine policy where the commercial distribution of content and the access to balanced information collide. Much of the problem, according to Halavais, can be attributed to a lack of policy transparency, creating a fertile ground for misuse and mistrust. As is the case with governments, he feels, informed citizenry and open discourse are crucial ingredients of effectiveness. As centralizing tendencies usually get reinforced over time, it seems fair to expect nothing less from the search engine industry. In light of the current debate, it seems fair to say that the idea of the 'seamless universe of the World Wide Web' as envisioned by Tim Berners-Lee upon its conception – designed in a decentralized manner, without central control and universally accessible – is presently unsustained. Introna and Nissenbaum therefore "urge engineers and scientists who adhere to the ideology of the Web, to its values of inclusivity, fairness, and scope of representation, and so forth, to pursue improvements in indexing, searching, accessing, and ranking with these values firmly in mind."[70]

In reference to the central question posed in the introduction to this thesis, the next chapter will describe the proposed plans for a Semantic Web. Although the concept of the Semantic Web does not comment on the search engine as such, but rather on the current state of data organization (or lack thereof) on the Web, It is important to note that Web search has developed in the last decade-and-a-half on top of a very dynamic and immensely scaled Web. Seeing how the ideologies that inspired the development of the original Web still play a huge role in current debates about Web search technology, we may ask to what extent these values are readdressed in the plans for augmentation.

[69] Alexander Halavais, Search Engine Society, Digital Media and Society Series (Cambridge: Polity Press, 2009).

[70] Introna and Nissenbaum, "Shaping the Web: Why the Politics of Search Engines Matters."

3 Semantics on the Web

3.1 Weaving a Semantic Web

In 2001 an article titled “The Semantic Web” appeared in the Scientific American, which officially introduced the concept to the academic world. Article authors were World Wide Web-inventor Tim Berners-Lee, James Hendler and Ora Lassila. The elaborate subtitle - “A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities” - already indicated the scope of the project, that by the time of this writing is still in development at the World Wide Web Consortium (W3C). The 2001 article proposes an extension of the mostly human and document oriented World Wide Web that will better enable human-computer interaction. By designing Web content in a formalized and structured way so that it is readable by computers and people alike, Berners-Lee et al. argue that software programs will be able to conduct *automated reasoning*; processing data in association with other data on the Web. The article clearly extends a vision for the future in which people create and instruct their own software agents to search and combine information from various online sources and to share this data between them.[71] An example on the W3C Semantic Web homepage illustrates this scenario:

> “The Semantic Web is a web of data. There is lots of data we all use every day, and it is not part of the web. I can see my bank statements on the web, and my photographs, and I can see my appointments in a calendar. But can I see my photos in a calendar to see what I was doing when I took them? Can I see bank statement lines in a calendar? Why not? Because we don't have a web of data. Because data is controlled by applications, and each application keeps it to itself.” [72]

Semantics, generally taken to mean the ‘study of meanings’[73], is used within the Semantic Web discourse - as well in the broader field of computer science - as a way to add structured data to objects within Web pages, so that the Web may be treated more like a relational database. The basis for the Semantic Web was already present at the time of Tim Berners Lee's 1989 proposal for the WWW. Employed by CERN[74] at the time, he wrote “Information management: A Proposal”, a paper in which he describes how relations can be indicated between people, technology and documents within a ‘data web’. In this proposal, a graph shows the relations between and the knowledge about these

[71] Tim Berners-Lee, James Hendler and Ora Lassila, "The Semantic Web," Scientific American 295.5 (2001).

[72] Ivan Herman, Semantic Web Activity, 2005, W3C <http://www.w3.org/2001/sw> July 21 2009.

[73] "Semantics." Merriam-Webster Online Dictionary, 2009, Merriam-Webster Online <http://www.merriam-webster.com/dictionary/semantics>

[74] Organisation Européenne pour la Recherche Nucléaire (European Organization for Nuclear Research)

objects; forming a Web of data he named 'Mesh' (see *figure 5*).[75]

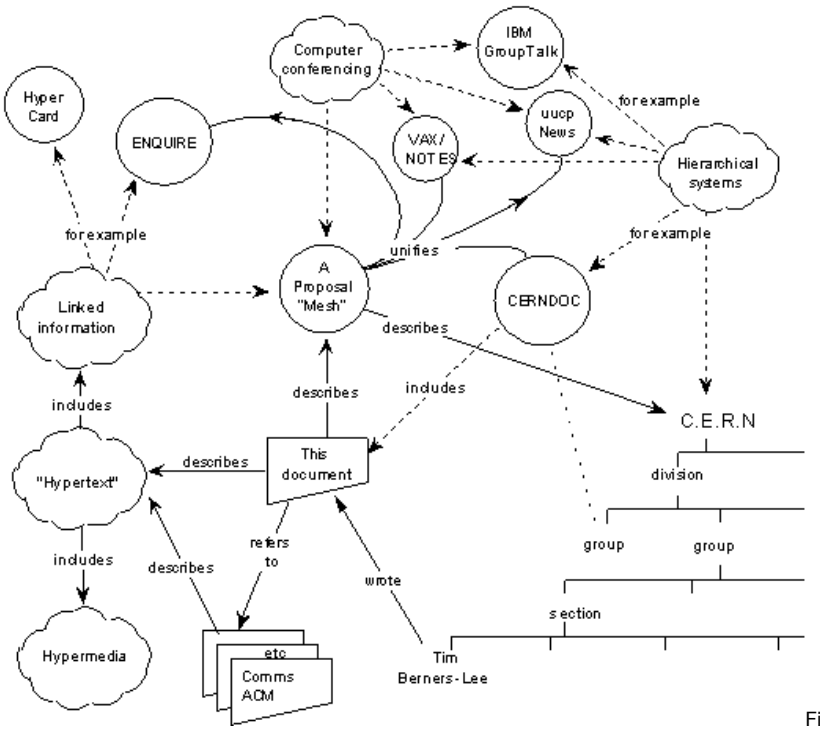

Figure 5. Mesh[76]

We can see some form of this reasoning on the Web today with the advent of the so-called semantic search engine. Semantic search engines attempt to disambiguate user intent and provide more relevant results by using a structure of denoted relations between data objects in a network. This technique is increasingly used - often in combination with natural language processing[77] - to search online collections of data in which formal relations between various objects are relatively easy to be defined. One example of a recent semantic search engine within a specific knowledge domain is Europeana (currently still in beta), a project that: "provides integrated access to digital objects from the cultural heritage organizations of all the nations of the European Union"[78]. The museums, archives and cultural organizations partnering with Europeana are asked to structure and annotate their databases according to a general, predefined set of terms, which makes their heterogeneous content uniformly searchable through the Europeana engine. The Europeana developers have established that within the domain of cultural heritage, user queries will most likely be based on "who, what, where and when" questions, to which the elements used for structuring the data (as exemplified in *figure 6*) correspond accordingly.[79]

[75] Tim Berners-Lee, Information Management: A Proposal, 1989, W3C <http://www.w3.org/History/1989/proposal.html> July 21 2009.

[76] Berners-Lee, Information Management: A Proposal.

[77] The ability to process queries that are formulated in human language, e.g. in the form of a question.

[78] Europeana, "Mapping & Normalization Guideline for Europeana Prototype." (2009).

[79] Europeana <http://www.europeana.eu/portal>

Strongly recommended	Recommended	Additional elements	Europeana elements
dc:title	dc:coverage	dc:format	europeana:country
dcterms:alternative	dcterms:spatial	dcterms:extent	europeana:hasObject
dc:creator	dcterms:temporal	dcterms:medium	europeana:isShownAt
dc:contributor	dc:description	dc:identifier	europeana:isShownBy
dc:date	dcterms:isPartOf	dc:rights	europeana:language
dcterms:created	dc:language	dcterms:provenance	europeana:object
dcterms:issued	dc:publisher	dc:relation	europeana:provider
	dc:source	dcterms:conformsto	europeana:type
	dc:subject	dcterms:hasFormat	europeana:unstored
	dc:type	dcterms:isFormatOf	europeana:uri
		dcterms:hasVersion	europeana:usertag
		dcterms:isVersionOf	europeana:year
		dcterms:hasPart	
		dcterms:isReferencedBy	
		dcterms:references	
		dcterms:isReplacedBy	
		dcterms:replaces	
		dcterms:isRequiredBy	
		dcterms:requires	
		dcterms:tableOfContents	

Figure 6. List of elements for the description of cultural objects in Europeana[80]

The idea behind semantic search is a formal one. In order for relations between data objects to be described, consensus must be in place about the nature of these relations. Much like the Wolfram|Alpha computational engine that relies on accepted mathematical models to define the structure of and relation between the available information in a data set, semantic search relies on a model that is given form by human experts within a particular domain of often culturally based knowledge. In a well-defined collection of information and with a limited amount of conflicting views as to the relations between objects, consensus may be reached eventually. Applying such formal structures to the global and diverse content of the World Wide Web is a different story.

3.2 Semantic Web Technology

In order to be able to discuss concerns raised about the Semantic Web ideals from a philosophical and Humanities perspective, and to address the concept of ontology that is crucial to doing so, it is needed to briefly dig into the technology that is involved with its development. The scope of this thesis and the complex nature of the technology at hand however urge me to limit this paragraph to discussing the two most important technologies that facilitate the Semantic Web: eXtensible Markup Language (XML) and the Resource Description Framework (RDF), which are both already developed

[80] Europeana, "Mapping & Normalization Guideline for Europeana Prototype."

and used.

XML is a language for creating tags to label Web documents or parts of text on a Web page. By using XML, webmasters can annotate and add structure to their data and documents. However, for the XML tags to be used by a software program for instance, a common framework is needed that defines which tag is used for what purpose. This framework is provided by RDF.[81] The RDF framework essentially allows people to make statements about resources. These statements are always expressed in so-called 'triples' of *subject*, *predicate* and *object*, as exemplified by *figure 7*. An example of such a triple would be: (Tim Berners-Lee) (is the inventor of) (The World Wide Web).

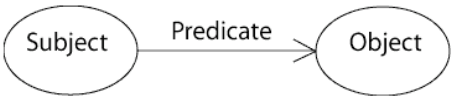

Figure 7. RDF triple[82]

Automatic reasoning is said to occur when computer programs apply 'inference rules' - rules that have an 'if-then' structure - to these statements. Taking several statements within structured data collections as premises, new information can be deduced automatically. For example: From the statements "trees are plants" and "oaks are trees", a software program may generate a third statement; "oaks are plants". This type of logic makes it possible to make assumptions on the basis of the metadata applied to objects and to generate information that was invisible before.

In order to allow anyone to define relations and to distinguish concepts from regular words or expressions in a Web page, the data objects as well as the relations between them are identified by so-called Universal Resource Identifiers (URIs). URIs are used as a means to name and access resources on the Internet. W3C states: "A resource should have an associated URI if another party might reasonably want to create a hypertext link to it, make or refute assertions about it, retrieve or cache a representation of it, include all or part of it by reference into another representation, annotate it, or perform other operations on it."[83] The most common type of URI known today is the Uniform Resource Locator (URL, as used in links on a Web page) that identifies a resource by the property 'network location'.

On a functioning Semantic Web as proposed by Berners-Lee et al., it should be possible for users to develop software programs or agents that collect online content from various databases and combine and share that data with software programs that are developed by other people. In order to do so effectively however, the two software programs, or rather the people behind them, must know that they are in fact talking about the same data. This is where the concept of ontology comes in. As is common to human language but not so to computer language, the word ontology has different meanings. In their article, Berners-Lee et al. point out that in the philosophical sense, ontology is "a

[81] Berners-Lee, Hendler and Lassila, "The Semantic Web."

[82] W3C, Resource Description Framework (Rdf): Concepts and Abstract Syntax, 2004 <http://www.w3.org/TR/rdf-concepts> July 21 2009.

[83] W3C, Architecture of the World Wide Web, Volume One, 2004, W3C <http://www.w3.org/TR/webarch> July 21 2009.

theory about the nature of existence, of what types of things exist." Within the field of artificial intelligence however, the term is taken to mean a "document or file that formally defines relations among terms."[84]

Ontologies in Artificial Intelligence or Web research are commonly used to resolve difficulties emerging around terminology. For instance, if in a database of addresses one field is called 'zip code', as common in the US, and in another database it is called 'postal code', as common in Europe, a conflict arises that makes it impossible for a software agent to automatically retrieve all relevant data at once.[85] An 'ontology' which states that postal code is in fact equivalent to zip code, is assumed to solve that problem. Several Semantic Web ontologies are already operational within specific knowledge domains. One example is 'FOAF' (derived from "friend of a friend"), an RDF based ontology that is used to semantically describe people and their social network. FOAF is aimed to be used within wikis for example, where it could serve to annotate user pages, or to describe articles about people. FOAF is sometimes referred to as a 'vocabulary', whose annotations may be imported into semantically managed Web pages.[86]

3.3 Ideals and Critique

Conflicts in terminology as described in the former paragraph are easier to resolve than conflicts that are based in a differing cultural or political understanding. In the philosophical take of the term, the idea of totally matching ontologies is a principal and political impossibility. This issue is addressed elaborately in an entry on the Nettime.org mailing list in September 2007 in which Florian Cramer, reader in Communication in a Digital Age at the Dutch Piet Zwart Institute, emphasizes the struggle between, or rather the mutual ignorance of both engineering and humanities that finds expression in the idea of the Semantic Web. He argues that, whereas computer scientists seem oblivious of the cultural load carried by terms such as 'ontology' and 'semantics', humanities researchers in turn seem unable to fully grasp the concept of the Semantic Web outside of a cultural understanding.

Cramer's critique focuses on the fact that universal ontologies – to which he actually prefers the term 'cosmologies' – have been designed all through the Middle Ages and the Renaissance. Complex tree-like classification schemes similar to those of the Semantic Web were used to structure human knowledge within encyclopedias before the arbitrary but revolutionary method of using the alphabet became the dominant paradigm. "Bluntly said", he argues, the Semantic Web is "nothing else but technocratic neo-scholasticism based on a naive if not dangerous belief that the world can be described according to a single and universally valid viewpoint; in other words, a blatant example of cybernetic control ideology and engineering blindness to ambiguity and cultural issues."[87] Clay Shirky makes a similar point, emphasizing the paradox that exists in the fact that the collective appliance of

[84] Berners-Lee, Hendler and Lassila, "The Semantic Web."

[85] Berners-Lee, Hendler and Lassila, "The Semantic Web."

[86] Semantic Web, Foaf, 2008 <http://semanticweb.org/wiki/FOAF> November 1 2009.

[87] Florian Cramer, <Nettime> Critique of The "Semantic Web", 2007, Nettime mailing list archives <http://www.nettime.org/Lists-Archives/nettime-l-0712/msg00043.html> July 21 2009.

meta data is problematic without the existence of a globally accepted ontology, while at the same time such a standard is fundamentally unreachable.[88]

The concepts of automatic reasoning and syllogism have been of philosophical significance since Aristotle, but may also make one think of Vannevar Bush's 1945 proposal for the Memex. In 'As We May Think', Bush pleads to make universal knowledge more accessible through the mechanical technology available at the time. Machines, he felt, could help people – scientists, in particular - reach a higher level of knowledge organization by taking over a great deal of logical and repetitive processes:

> "Whenever logical processes of thought are employed—that is, whenever thought for a time runs along an accepted groove—there is an opportunity for the machine. Formal logic used to be a keen instrument in the hands of the teacher in his trying of students' souls. It is readily possible to construct a machine which will manipulate premises in accordance with formal logic, simply by the clever use of relay circuits. Put a set of premises into such a device and turn the crank, and it will readily pass out conclusion after conclusion, all in accordance with logical law, and with no more slips than would be expected of a keyboard adding machine."[89]

The article, often thought to envision the hypertext system, may as well be said to envision the Semantic Web ideal as Bush describes the way numerous documents may be linked together, forming trails and side-trails of documents connected by code words. The Memex however, being a personal desk-like mechanical device, was not a networked technology and the linkage of documents depended on the logic of its owner exclusively. Obviously, this is the neither the case for the WWW nor for the Semantic Web and the annotation of documents available online today. Clay Shirky suggests that Semantic Web statements serve as input for inference rules not because this logic is a great way to handle subjective and context-dependent relations – he feels it is not – but rather because this logic is a thing that computers handle well.[90]

Both Cramer and Shirky strongly oppose the ideals of the Artificial Intelligence project which they feel are guided by a technological imaginary that is seemingly blind to the many historical attempts – and their imminent failure – to capture the world in systems of sorts. It should be mentioned that Berners-Lee and his co-authors do not refrain to reflect on the fact that their technology is bound to encompass paradoxes and questions that can't be answered. Expressing much the same value set that accompanied the development of the WWW, they name 'universality' and the fact that 'anything may link to anything' as the core features of the Semantic Web. The authors promote the breadth of information, noting that there should be no discrimination between different kinds of data, be it cultural, academic or 'scribbled draft'. Furthermore, Semantic Web researchers reject the centralized model of traditional knowledge representation systems and adhere to decentralization as the guiding principle for further developments. In the larger quest for versatility then, conflicts and paradoxes are

[88] Clay Shirky, "The Semantic Web, Syllogism, and Worldview," Clay Shirky's Writings About the Internet (2003), vol. 2009.

[89] Bush, "As We May Think."

[90] Shirky, "The Semantic Web, Syllogism, and Worldview."

considered facts to accept rather than to fight against. Much like the early stages of the Web's development, the authors argue, oppositional voices can be heard that question the effectiveness of the new medium without any central tree-structured database, but the articulate power of the technology has allowed for a vast amount of information to become available, and for search engines to be built upon it and make it all findable.[91]

These are strong values expressed in 2001 as the technology was introduced to the academic world. Currently, the many pages of the W3C Semantic Activity website exclusively deal with the broad range of scripting and programming languages that have been developed, the way in which these may be put to use and the practical advantages of doing so. Interestingly, there is no reference to be found to the debate that has taken shape in the decade following the Scientific American article – neither from the perspective of scholars such as Cramer and Shirky nor from the critique surrounding Web search in general and Google in particular. It has long become clear that the decentralizing, inclusive qualities of the Web propagated by its developers are not sustained today. Net neutrality issues have pointed out that there is a strong economic incentive to challenge the values by which the Web was conceived. Moreover, as the Web now practically equals Google search results, we can clearly see the centralized model of traditional knowledge representation systems reappear. 'Making information more accessible' seems to have become the core goal of the W3C project, and we may consider its silence surrounding the matter of politics a waive of responsibility. Listening closely though, it does tell us something: "Don't blame us, the computers did it".

Now and then, individual voices such as that of Steven Pemberton, a CWI[92] researcher involved in various W3C commissions, speak out about the danger of 'walled gardens' on the Web, touching part of the discussion on privacy and the power of search giants such as Google. In two recent conference lectures, Pemberton sketches an image of the current Web as being divided into several sub webs by software programs that proprietize data.[93] The strength of the network as a whole, he feels, is compromised by 'Web 2.0' applications that rely on users to add value and commit themselves, but at the same time confront them with privacy issues and the risk of losing data when such a site closes down or changes policy. Taking the example of a Google account, this would imply one losing emails, agenda information, documents and a host of other data as services rapidly expand. Instead, he urges users to keep their RDF annotated data on their own personal sites and share it only selectively through the use of APIs.[94] Interestingly, Pemberton, who was not speaking on behalf of the W3C, directs his argumentation toward the individual user instead of the developer community and focuses on one core technology (RDFa) instead of the entire Semantic Web concept.

[91] Berners-Lee, Hendler and Lassila, "The Semantic Web."

[92] Centrum Wiskunde & Informatica, the national research center for mathematics and computer science in the Netherlands.

[93] The NLUUG conference 'The Open Web' in Ede, the Netherlands on October 29 2009, and the 'Society of the Query' conference in Amsterdam, the Netherlands on November 13-14 2009.

[94] Jos Poortvliet, "Walled Gardens, Semantic Data and the Open Web: an Interview with Steven Pemberton," KDE.News (2009), vol. 2009.

In my opinion, this practice foregrounds some profound contradictions that lie at the base of the Semantic Web development process and its institutionalization within the W3C. The final part of this thesis will elaborate on these contradictions and address the feasibility of a Semantic Web coming into existence in the near future.

3.4 The Future of the Web

In light of the absence of a critical debate surrounding Semantic Web developments, it is important to consider an argument outlined by Alexander Galloway in his 2004 book *Protocol*. The lack of any centralized administration or control on the Internet, Galloway states, does in no way imply the absence of control as such. Protocological standards such as HTML strongly embody a form of control in their definition of the possibilities and impossibilities of the actions that may be performed over and through it. In both the WWW as well as Semantic Web developments the creators have strongly adhered to building a distributed architecture, but Galloway stresses that all protocols actively shape the boundaries and conditions for technological innovation with the aim to standardize and organize. The social utopia of openness and interactivity sketched by computer scientists such as Tim Berners-Lee is possible *only* through radical homogenization and broad adoption of technological standards, or in Galloway's words: "In order to be politically progressive, protocol must be partially reactionary".[95] This 'generative contradiction' has been part of the Web since its earliest beginnings, he argues, and is additionally reflected in the fact that in order to be able to develop and implement standards aimed at achieving a distributed architecture, peer groups of scientists must be organized into cumbersome, adistributed bureaucracies. Galloway calls this institutionalization a type of 'tactical standardization'.[96] In this sense, we may argue that it would in fact be counterintuitive for the standardizing body of the W3C to focus on the diversity of user concerns and the possibility of overcoming these by partially implementing Semantic Web technologies. After all, the open and distributed nature of the future Web relies fully on a homogeneous adaption to standards.

There is a final, pragmatic issue I would like to address that arises when thinking about the feasibility of a future Semantic Web. Leading the emphasis away from both the ideological or epistemological debate and the proposed advantages of a functioning Semantic Web, and towards the factual realization process, we may take a better look at the effort required of the Web developer community and the authority of the W3C within this process. I will do so by using the example of a new HTML standard that has been under development at W3C since 2004.

HTML5, the newest version of the core Web mark-up language, aims to bring an increased level of structure and semantics to websites. Widespread implementation of the new standard would signify an important step toward structured content.[97] Functioning as a follow-up to HTML4.01 and XHTML1.0, HTML5 adds new elements to HTML for the first time in almost a decade.[98] While the new

[95] Alexander Galloway, Protocol (Cambridge: The MIT Press, 2004). 142

[96] Alexander Galloway, Protocol 138-143

[97] WHATWG, Faq, 2009 <http://wiki.whatwg.org/wiki/FAQ> November 1 2009.

[98] Elliotte Rusty Harold, "New Elements in Html 5," IBM DeveloperWorks (2007), vol. 2009.

standard is aimed to be partially implementable in late 2010 (most HTML versions under development pre-issue in 'transitional' mode) it is suggested that the standard will not reach the stage of becoming a so-called W3C recommendation before the year of 2022. A lengthy time frame, but not at all uncommon within the HTML development process since a vast amount of tests precede any official recommendation.[99] An interesting study by Beatty et al. in 2008 however, notes that due to the lack of authority of W3C to enforce the adoption of new standards, and also due to the wide range of HTML versions available, Web developers most of the time reside to using flexible, transitional versions of a standard that offer some of the new features but are significantly less stringent and easier to implement. Especially when formatting requirements become tighter – as would be the case with HTML5 – the inclination toward a more permanent use of a transitional version of a standard is strong. Beatty et al. support this thesis with a practical survey of the Alexa.com top 100.000 most popular websites, in which the researchers found that less than 2.5% of the sites validate as being strict HTML versions. 22% of the surveyed sites did not offer the mandatory DOCTYPE declaration (so no statement could be made) and 75.5% of the most visited sites used transitional versions.[100]

These are shocking figures, and form quite a contrast to the giant scope of webmasters' compliance to Google's SEO standards. The remarkable difference here may be argued to exist in the social pressure that is exerted by Google through its sanctions of removing a site from their index and as such rendering it invisible. While both parties enjoy no 'official' authority, there is a significant and urgent benefit to complying with Google's standards that seems to lack on the part of W3C. As far as the adoption of Web standards goes, the market is not regulated and there are neither rules for minimum website quality nor sanctions for non-compliance. An additional contradiction, then, to the inherent power of protocol to shape values on the Web, is the complete lack of power of the W3C to secure the necessary level of implementation and as such secure the values to come into being. This being the case, it will be solely up to the benefits of the new standard to outweigh objections to its implementation. Only then does it stand a chance of ever becoming widely accepted and, as such, actually useful.

[99] WHATWG, Faq.

[100] Patricia Beatty, Scott Dick and James Miller, "Is HTML in a Race to the Bottom? A Large-Scale Survey and Analysis of Conformance to W3C Standards," IEEE Internet Computing 12.2 (2008).

Conclusion

We have seen Web search go through many stages of development throughout the past two decades. The emergence of full-text search in the mid-nineties has caused the search engine to become the focal point of the Web search process. The subsequent move from a quantative focus on indexing toward a qualitative focus on relevance ranking has rendered the engine with even more authority, and sparked the start of a critical debate about the politics of Web search by the beginning of this century. Even today, with the amount of critical theory on the subject steadily growing, we are trying to put foothold to the discussion. As it has originated and taken shape largely in the United States, a European debate is initiated slowly but surely over the course of the past two years and has yet to reach and involve the broad user community. During the recent 'Society of the Query' conference in Amsterdam, Siva Vaidhyanathan aptly noted how in order to even be aware of the concerns that search engines raise, people must first care. Currently, he argues, "only the elite and the proficient get to opt out".[101]

In the midst of establishing a strong and critical Web search debate, an augmentation of the World Wide Web is under development at W3C, with the aim of building a decentralized and inclusive Web architecture in which relations between data objects are semantically defined. Semantic Web developers are quick to 'label' this future Web with much the same value set that they invested in the development of the original WWW. Seeing this it is very important to note that the Web is no longer new; any plans for a 'second coming' should take into account the way in which the contemporary Web, through search engine technology, appears neither inclusive nor decentralized. We may seriously question the way in which such values are repeated without paying reference to a decade of critical thought. In the same line of thinking, it is astonishing that there is hardly any scholarly work to be found that focuses on the topic of the Semantic Web and takes into account the ethical, political, philosophical and cultural facets of such an endeavor. The overall non-existence of a lively and balanced debate on Semantic Web developments is problematic, to say the least.

The very abstract phenomenon of the Semantic Web is contrasted by the extreme service-mindedness of today's search engine; striving for ease of use and aiming at long-term commitment and trust. As a profoundly 'social engine', Google derives its power from a deep understanding of the workings of the Web and the distribution of attention. In the daily reality of Web use, then, the Semantic Web is no more than a phantom phrase that one may accidentally come across, invoking, perhaps, a faint idea about computers understanding natural language. As much as the Artificial Intelligence project may want and need it to be, the world in reality is not a relational database. It is exactly this quality that distinguishes computers from people, and that distinguishes the dream of a Semantic Web from its rather 'syntactic' reality.

[101] Chris Castiglione, "Siva Vaidhyanathan on Googlization, "Only the elite and proficient get to opt out"," Society of the Query (2009), vol. 2009.

The Web has grown explosively over the years, not merely in the amount of Web pages but at least as much in terms of cultural diversity, geography and economic value. Questions may be raised as to the power of protocol and large bureaucratic institutions such as W3C to initiate reformations as vast as the Semantic Web. As argued in former chapters, W3C cannot make explicit political demands, it wields no social power and is inherently focused on the broad adaption of standards rather than individual concerns and benefits. Considering the above, and as protocol is only as strong as its implementation, it may be argued that the Web has become too large for protocol to function as an ‘engine of change’ any longer.

References

Alpert, Jesse, and Nissan Hajai. "We knew the web was big..." The Official Google Blog 2008. Vol. 2009. <http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html>

Anderson, Chris. "The Long Tail." Wired 2004. <http://www.wired.com/wired/archive/12.10/tail.html>

Arms, William Y. Digital Libraries. Cambridge: MIT Press, 2001.

Barabási, Albert-László. Linked: How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life. New York: Plume Books, 2003.

Barabási, Albert-László, and Réka Albert. "Emergence of Scaling in Random Networks." Science 286.5439 (1999): 509-12.

Battelle, John. "Google Announces New Index Size, Shifts Focus from Counting." John Battelle's Searchblog 2005. Vol. 2009. <http://battellemedia.com/archives/001889.php>

---. "In This Battle, Size Does Matter: Google Responds to Yahoo Index Claims." John Battelle's Searchblog 2005. Vol. 2009. <http://battellemedia.com/archives/001790.php>

BBC NEWS. "New Google 'puts Bing in shade'". 2009. (August 12). October 23 2009. <http://news.bbc.co.uk/2/hi/technology/8195739.stm>.

Beatty, Patricia, Scott Dick, and James Miller. "Is HTML in a Race to the Bottom? A Large-Scale Survey and Analysis of Conformance to W3C Standards." IEEE Internet Computing 12.2 (2008): 76-80.

Berkeley, UC. "Invisible or Deep Web: What it is, How to find it, and Its inherent ambiguity". 2009. October 23 2009. <http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/InvisibleWeb.html>.

Berkhin, Pavel. User-sensitive pagerank. YAHOO! INC. assignee. 2008.

Berners-Lee, Tim. "Information Management: A Proposal". 1989. W3C. July 21 2009. <http://www.w3.org/History/1989/proposal.html>.

---. "WorldWide Web Seminar". 1991. July 27 2009. <http://www.w3.org/Talks/General.html>.

Berners-Lee, Tim, James Hendler, and Ora Lassila. "The Semantic Web." Scientific American 295.5 (2001): 34-43.

Bleiweiss, Evan. "A Brief History of Meta Tags." Media\Outreach: Media Outreach, 2008. Vol. 2009. <http://mediaoutreach.com/2008/11/a-brief-history-of-meta-tags/>

Boulton, Clint. "Google Search Ranking Feature Threatens Wikia Search." eWeek 2008. Vol. 2009. http://www.eweek.com/c/a/Search-Engines/Google-Search-Ranking-Feature-Threatens-Wikia-Search/

Bush, Vannevar. "As We May Think." The Atlantic (1945). <http://www.theatlantic.com/doc/194507/bush>

Carr, Nicholas. "The Centripetal Web." Rough Type 2008. Vol. 2009. <http://www.roughtype.com/archives/2008/10/the_centripetal.php>

Castiglione, Chris. “Siva Vaidhyanathan on Googlization, “Only the elite and proficient get to opt out”,” Society of the Query 2009. Vol. 2009. <http://networkcultures.org/wpmu/query/tag/siva-

vaidhyanathan/>

Cornett, Larry. "Search & Serendipity: Finding More When You Know Less." Search Engine Land 2007. Vol. 2009. <http://searchengineland.com/search-serendipity-finding-more-when-you-know-less-11971>

Cramer, Florian. "<nettime> Critique of the "Semantic Web"". 2007. Nettime mailing list archives. July 21 2009. <http://www.nettime.org/Lists-Archives/nettime-l-0712/msg00043.html>.

Cutts, Matt. "How does Google collect and rank results?". 2009. Google. July 21 2009. <http://www.google.com/librariancenter/articles/0512_01.html>.

---. "Minty Fresh Indexing." Matt Cutts: Gadgets, Google, and SEO 2007. Vol. 2009. <http://www.mattcutts.com/blog/minty-fresh-indexing/>

Digital Methods Initiative. "The Demise of the Directory: Web librarian work removed in Google". 2008. October 23 2009. <http://wiki.digitalmethods.net/Dmi/DemiseDirectory>.

DMOZ. "About the Open Directory Project". 2009. October 23 2009. <http://www.dmoz.org/about.html>.

"Europeana". <http://www.europeana.eu/portal/>.

Europeana. "Mapping & Normalization Guideline for Europeana Prototype." 2009. <http://www.version1.europeana.eu/c/document_library/get_file?uuid=104614b7-1ef3-4313-9578-59da844e732f&groupId=10602>

Galloway, Alexander. Protocol. Cambridge: The MIT Press, 2004.

Google. "Corporate Information: Technology Overview". 2009. Google. July 21 2009. <http://www.google.com/corporate/tech.html>.

Google Web Search. "Features: SearchWiki". 2009. October 23 2009. <http://www.google.com/support/websearch/bin/answer.py?hl=en&answer=115764>.

Google Webmaster Central. "Search Engine Optimization (SEO)". 2009. October 23 2009. <http://www.google.com/support/webmasters/bin/answer.py?answer=35291&cbid=1qcfyilhw8zwr&src=cb&lev=answer>.

Graham, Paul. "Metatags -- The Latest Developments." Computer Fraud & Security 2000.10 (2000): 12-13.

Grehan, Mike. "Search, That Was Mighty Sociable." Clickz: Incisive Interactive Marketing LLC, 2008. Vol. 2009. <http://www.clickz.com/3630618>

Grimmelmann, James. "The Google Dilemma." New York Law School Review 53.939 (2009).

Halavais, Alexander. Search Engine Society. Digital Media and Society Series. Cambridge: Polity Press, 2009.

Harold, Elliotte Rusty. "New elements in HTML 5." IBM DeveloperWorks 2007. Vol. 2009. <http://www.ibm.com/developerworks/library/x-html5/>

Herman, Ivan. "Semantic Web Activity". 2005. W3C. July 21 2009. <http://www.w3.org/2001/sw>.

---. "W3C Semantic Web Activity". 2009. October 27 2009. <http://www.w3.org/2001/sw/>.

Huang, Chun-Yao, et al. "Concentration of Web users' online information behaviour." Information Research 12.4 (2007).

Internet Archive, The. "http://www.google.com September 25". 2005. Internet Archive Wayback

Machine. July 21 2009. <http://web.archive.org/web/20050924172505/www.google.com>.

---. "http://www.google.com September 26". 2005. Internet Archive Wayback Machine. July 22 2009. <http://web.archive.org/web/20050926235645/www.google.com>.

Introna, Lucas D., and Helen Nissenbaum. "Shaping the Web: Why the Politics of Search Engines Matters." The Information Society: An International Journal 16.3 (2000): 169 - 85.

Jeanneney, Jean-Noël. Google and the Myth of Universal Knowledge. Chicago: University of Chicago Press, 2007.

Kaiser, Cameron. "Down the Gopher Hole." TidBits 2007. Vol. 2009. <http://db.tidbits.com/article/8909>

Kleinz, Torsten. "Wikia Search opens its doors." The H 2008. Vol. 2009. <http://www.h-online.com/newsticker/news/item/Wikia-Search-opens-its-doors-735391.html>

Mei, Qiaozhu, and Kenneth Church. "Entropy of search logs: how hard is search? with personalization? with backoff?" Proceedings of the international conference on Web search and web data mining. ACM.

Microsoft. "About Bing". 2009. Microsoft Corporation. July 21 2009. <http://www.discoverbing.com/behindbing/about.aspx>.

---. "Bing". 2009. Microsoft. July 21 2009. <http://www.bing.com>.

Netcraft. "September 2009 Web Server Survey". 2009. October 23 2009. <http://news.netcraft.com/archives/2009/09/23/september_2009_web_server_survey.html>.

Poortvliet, Jos. "Walled Gardens, Semantic Data and the Open Web: an Interview with Steven Pemberton." KDE.News 2009. Vol. 2009.

Radu, Remus, and Raluca Tanase. "Lecture #4: HITS Algorithm - Hubs and Authorities on the Internet". 2009. Cornell University. July 21 2009. <http://www.math.cornell.edu/~mec/Winter2009/RalucaRemus/Lecture4/lecture4.html>.

Schonfeld, Erick. "Wowd Takes A Stab At Realtime Search With A Peer-To-Peer Approach." TechCrunch 2009. Vol. 2009.

Search Engine Laisha's List. "Feature: ODP History". 1999. October 23 2009. <http://www.laisha.com/zine/odphistory.html>.

Semantic Web. "FOAF". 2008. November 1 2009. <http://semanticweb.org/wiki/FOAF>.

Shirky, Clay. "The Semantic Web, Syllogism, and Worldview " Clay Shirky's Writings About the Internet 2003. Vol. 2009.

StatCounter Global Stats. "Top 5 Search Engines from Sep 08 to Oct 09". 2009. October 23 2009. <http://gs.statcounter.com/#search_engine-ww-monthly-200809-200910>.

Tabke, Brett. "A Brief History of SEO: The real search engine wars." WebmasterWorld 2002. Vol. 2009. <http://www.webmasterworld.com/forum5/1008.htm>

"The WWW Virtual Library". October 10 2009. <http://vlib.org/>.

Times, The New York. "Yahoo-Microsoft Deal". 2009. (July 30). October 23 2009. <http://topics.nytimes.com/top/news/business/companies/yahoo_inc/yahoo-microsoft-deal/index.html>.

Treese, Win. "Is the power of web search diminishing?" netWorker 12.2 (2008): 13-15.

Underwood, Lee. "A Brief History of Search Engines". 2004. October 23 2009. <http://www.webreference.com/authoring/search_history/>.

Vaidhyanathan, Siva. "Another Chapter: the many voices of Google." The Googlization of Everything 2009. Vol. 2009. <http://www.googlizationofeverything.com/2009/06/another_chapter_the_many_voice.php>

W3C. "Architecture of the World Wide Web, Volume One". 2004. W3C. July 21 2009. <http://www.w3.org/TR/webarch>.

---. "Resource Description Framework (RDF): Concepts and Abstract Syntax". 2004. W3C Recommendations. Ed. Brian McBride. July 21 2009. <http://www.w3.org/TR/rdf-concepts>.

Wales, Jimmy. "Update on Wikia – doing more of what's working." Jimmy Wales 2009. Vol. 2009. <http://blog.jimmywales.com/2009/03/31/update-on-wikia/>

Wall, Aaron. "History of Search Engines: From 1945 to Google 2007". 2007. October 23 2009. <http://www.searchenginehistory.com/>.

Web Search Workshop, The "AltaVista: A brief history of the AltaVista search engine". 2009. The Web Marketing Workshop Ltd. July 20 2009. <http://www.websearchworkshop.co.uk/altavista_history.php>.

WHATWG. "FAQ". 2009. November 1 2009. <http://wiki.whatwg.org/wiki/FAQ>.

Wolfram|Alpha. "About Wolfram|Alpha". 2009. October 23 2009. <http://www.wolframalpha.com/about.html>.

---. "Frequently Asked Questions". 2009. October 23 2009. <http://www.wolframalpha.com/faqs.html>.

Yahoo! "Yahoo! Directory". Website. Yahoo! Inc. July 20 2009. <http://dir.yahoo.com>.