Search engines and the production of academic knowledge

van Dijck, J.

DOI
10.1177/1367877910376582

Publication date
2010

Document Version
Submitted manuscript

Published in
International Journal of Cultural Studies

Citation for published version (APA):
van Dijck, J. (2010). Search engines and the production of academic knowledge. *International Journal of Cultural Studies*, *13*(6), 574-592. https://doi.org/10.1177/1367877910376582

INTERNATIONAL
journal *of*
CULTURAL studies

© The Author(s), 2010.
Reprints and permissions:
http://www.sagepub.co.uk/journalsPermissions.nav
www.sagepublications.com
Volume 13(6): 574–592
DOI: 10.1177/1367877910376582

Search engines and the production of academic knowledge

● José van Dijck
University of Amsterdam, The Netherlands

ABSTRACT ● This article argues that search engines in general, and Google Scholar in particular, have become significant co-producers of academic knowledge. Knowledge is not simply conveyed to users, but is co-produced by search engines' ranking systems and profiling systems, none of which are open to the rules of transparency, relevance and privacy in a manner known from library scholarship in the public domain. Inexperienced users tend to trust proprietary engines as neutral mediators of knowledge and are commonly ignorant of how meta-data enable engine operators to interpret collective profiles of groups of searchers. Theorizing search engines as nodal points in networks of distributed power, based on the notions of Manuel Castells, this article urges for an enriched form of information literacy to include a basic understanding of the economic, political and socio-cultural dimensions of search engines. Without a basic understanding of network architecture, the dynamics of network connections and their intersections, it is hard to grasp the social, legal, cultural and economic implications of search engines. ●

KEYWORDS ● e-literacy ● Google Scholar ● Google Search ● information literacy ● knowledge production ● power/knowledge ● profiling systems ● ranking systems ● search engines

In the past decade, search engines have become indispensable tools in the construction of scholarly knowledge. Digitized search has evidently changed the way we learn and read, and it might well be argued that the production of scholarly knowledge has never been easier because we now have more access to more sources than ever before (Carr, 2008a). A college student assigned to write a paper or thesis these days, is likely to start with Google Search or Google Scholar. When you ask students how they researched their topic, their likely answer is that they 'googled' it, meaning they clicked on the 10 top-ranked results. The role of libraries and librarians has changed dramatically from mediators in the process of information searching to facilitators of digital processes. One reason why students appreciate the university library these days is for offering LibraryLink, a Google service which, coupled to the library's online journal subscriptions, provides convenient access to full-text downloads. As information specialist Stephanie Taylor (2007: 5) points out, many students think of library services as 'an add-on to the Google Scholar service, rather than the other way around'. Search engines have become dominant instruments in the production of knowledge and they are commonly regarded as neutral tools for information gathering.

This article argues that search engines in general, and Google Scholar in particular, are co-producers of academic knowledge. Knowledge is not simply conveyed to users, but is co-produced by search engines' ranking systems and profiling systems, none of which are open to the rules of transparency, relevance and privacy in a manner known from library scholarship in the public domain. Production of knowledge, as we have learned since Foucault (1980), is intricately intertwined with power and the mechanisms of power distribution. In order to identify how knowledge gets produced, we cannot simply examine search engines as objects; instead we need to regard them as 'actor networks' (Latour, 2005) – human–technological systems which are key to digitized knowledge production. Analyzing human–technological networks and the way they frame users' skills is part of the activity that Castells (2009: 431) calls the 'unwiring' of network activity. Following Castells' larger project to reveal the complex power relationships of digital networks, this article proposes to 'unwire' the construction of academic knowledge through the coded dynamics of search engines.

A critical discussion of digital knowledge production is important because society needs students and scholars who are not only competent users of search engines, but who can also reflect critically on the principles of knowledge formation. Teaching information literacy, including the strengths and weaknesses of various proprietary and public search tools, may partly remedy the growing influence of web search engines. If information literacy is restricted to the mere technicalities of how and when to use web-based search tools, however, students will miss out on a crucial reflective dimension. Knowledge is not simply brokered ('brought to you') by Google or other search engines, but is the result of intricate interactions between search engines and users' behaviour, channelled by the architecture and politics of network distribution. Students and scholars need to grasp the implications of these mechanisms in order to understand thoroughly the extent of networked power. Therefore, I propose to expand information literacy to include reflective skills on the social construction of knowledge, and thus to account for the political and ideological dimensions of automated search.

How does Google Scholar work (best)?

For decades, library information and reference systems have relied on transparent yet complex systems for indexing. As a student, finding your way amid the giant numbers of academic sources often required the help of professional librarians trained in the coded structure of the reference labyrinth. The Dewey system, which relies on appropriate labelling of keywords, never was the easiest system to use, and even if electronic library services (such as Web of Science, Metalib, Project Muse, etc.) made searching faster and more efficient, users' dependence on key words applied by reference librarians and publishers restricted retrieval and slowed down access. University libraries serve academic interest by filtering, ordering and ranking quality materials on the basis of their evaluated academic weight. They have always been public service institutions, relying on traditional library values, such as usefulness and reliability, neutrality and transparency, independence and respecting the right to privacy and confidentiality of their users (Caufield, 2005). These values are profound to the constitution of scientific knowledge. The obvious question thus arises: *how does Google Scholar compare to these library values firmly rooted in the public domain?*

Since Google Scholar's arrival in 2004, academics and librarians have extensively discussed the engine's effectiveness and usefulness, deliberated its reliability and relevance, and evaluated its merits and shortcomings vis-a-vis library-based web tools. As far as usefulness is concerned, it appears that the proprietary engine is extremely effective in finding precise quotes, citations or specific authors. Library scientist Bruce White agrees that Google Scholar 'works best for very tightly defined searches, those for which conventional database searches produce few or no results' (2007: 22). Especially now that full text copies of articles are increasingly accessible on the Internet – if not locked behind a publisher's paid access portal – it has become easier to trace exact quotations and references back to their original publication. While the bèta-versions of Google Scholar still suffered from overblown citation counts on its results pages – for instance, bogus author names such as 'F. Password' resulted in thousands of ranked hits – these retrieval problems have been more or less countered in recent versions of the engine (Jacso, 2008). Over the six years of its development, Google Scholar has improved its functionality when it comes to tracking well defined, precise scientific sources.

When Google Scholar arrived in 2004, it was applauded as a system that was better in terms of accessibility and comprehensiveness than most existing library systems combined – an electronic aid that would soon replace many traditional library functions (Gorman, 2006); it was also highly recommended for its democratizing potential, certainly in parts of the world that do not have access to well-resourced libraries. In brokering between information and knowledge, Google Scholar was expected to be the reliable, neutral and transparent (re)search tool the academic community had been waiting for – a kind of super-reference librarian that automatically extracts citations from reference lists and databases, and recognizes scientific documents. Indeed, Google Scholar promises to be a service that helps you identify the most relevant research across the world of scholarly research by adopting the standards of scientists and librarians. As we can read in its manual:

> Google Scholar aims to sort articles the way researchers do, *weighing* the full text of each article, the author, the publication in which the article appears, and *how often* the piece has been cited in other scholarly literature. The most *relevant* results will always appear on the first page.[1]

In its stated aims and function, Google Scholar replicates some scientific core values (such as citation analysis and weighed evaluation) and assumes some of the precepts that guide librarians in their work (e.g. selection for relevance).

At first sight, Google Scholar adopts one of the basic academic values – citation analysis – by letting algorithmic web spiders create indexes to a vast web of academic materials. Like its parent engine, Google Scholar functions as a ranking system based on semantic links to a vast reservoir of sources that through their provenance might be considered academically sound. However, Google Scholar's algorithm works on the basis of *quantitative* citation analysis, a process different from the one scholars use in their protected academic universe, where citations are also scored according to their relative status and weight in their specific professional disciplines. Ranking information through Google Scholar is quite similar to Google Search in that it ranks sources on the basis of *popularity* rather than truth-value or relevance. In the Scholar context, there is no clear peer-review system or citation analysis system that publicly lays out its ranking principles; there is only an algorithm – PageRank, named after its inventor Larry Page – that takes the number of links and hits as its basic units of ranking, but whose exact working is a well-kept trade secret. PageRank is a quantitative rather than qualitative system: a source that is well linked to other sources and is often clicked on thus gains in ranking, regardless of the document's status, relevance or value. As library scholar Margaret Markland (2005: 25) observes: 'Google equates "linking to a page" as "assigning importance," but this definition of importance may not necessarily indicate quality.'

Another core value in the public library reference system is transparency in terms of its covered sources. There have been a number of studies comparing Google Scholar's comprehensiveness to other (public and commercial) web-based services, such as Metalib, Scopus, Web of Science, Ingenta, Muse, etc.[2] In general, these studies point to a few profound principles where Google Scholar's lack of public values is especially apparent. First, they point at Google's incomplete and indeterminate coverage of source material. Despite the engine's crawl in a vast reservoir of published and unpublished materials, Google Scholar's coverage of scientific sources is notoriously incomplete because a number of scholarly societies (such as the American Chemical Society) or important publishers (such as Elsevier) refuse to give access to their databases. To properly evaluate Google Scholar's scope, one would need a precise list of databases covered by the engine, but so far Google has never published a list of scientific journals crawled by its spiders. Furthermore, Google Scholar shows a considerable time lag in citing newly published items, as the engine is not as regularly updated as conventional scholarly databases (Neuhaus et al., 2006). Google does not reveal the frequency of its updates, and neither does it provide time stamps for its ranked results, so a researcher can never trace the 'history' of a document (Hellsten et al., 2006). In other words, users do not know *how* the engine's coverage is limited in terms of scope and time span.

One could argue that Google Scholar covers a much larger number of sources than university libraries do in their web databases, but here, again, the public value of sorting relevant and weighed items enters the equation. Alongside authorized copies of peer-reviewed published articles Google Scholar also refers to various kinds of unofficial or grey sources, such as working papers, unpublished material, documents in preprint repositories, PowerPoint presentations published on university web sites and lecture notes. When academics decide to publish articles on the web that have not yet been peer-reviewed, one could argue, research results are accessible and retrievable much earlier than their official publication date. As valuable as these sources may turn out to be, their uncertain status and undefined quality make it hard to gauge these documents' value, especially for inexperienced college students. In addition, sources are often undefined and, consequently, their publishing context gets lost in the presentation of the full text. Dubious hits will appear legitimate when shown within an interface branded as scholarly, and, even if users are aware of these limitations, this might be misleading (Taylor, 2007). The indiscriminate ranking of sources on a page obscures the importance of the document's value based on its academic status. In other words, disclosure of the paper's status is entirely left to authors and its evaluation is entirely up to the user. The question is, of course, whether students can discern these variations of quality in listed documents.

Unfortunately, there is as yet little empirical or ethnographic research data illustrating how students actually go about open searches; however, surveys prove that students performing topic searches for scholarly papers overwhelmingly choose search engines, rather than library-offered research discovery networks, as their preferred starting-point.[3] Google Search – and, by extension, Google Scholar – have come to predominate over all other electronic information services offered by libraries. Electronic library services are less known about than search engines. In contrast to search engines, electronic library systems offer filtered sources that are carefully selected, indexed and described by specialists; their function as gatekeepers in the process of knowledge formation is based on public (library) values such as quality assessment, weighed evaluation and transparency in listed source databases. Students' search behaviour shows a preference for speed and convenience at the expense of quality and weighed relevance (Markland, 2005: 25–9). Although there are considerable differences in ways people make use of search engines, inexperienced users can be expected to have a strong penchant towards automated desktop queries, and generally resort to the most convenient tool.

Less than six years after their arrival, search engines – Google Scholar most prominently among them – have imposed a layered technological network on what used to be a library-based system of indexing and referencing sources of knowledge, and have come to dominate user patterns of information search. Automated search systems developed by commercial Internet giants like Google tap into public values scaffolding the library system and yet, when looking beneath this surface, core values such as transparency and openness are hard to find. The powerful technological network that increasingly filters access to (public) knowledge is itself rooted in obscure or simply unknown principles. Castells (2009: 45) has defined this 'network strategy' as 'the ability to connect and ensure the cooperation of different networks by sharing common goals and combining resources, while fending off competition from other networks by setting up strategic cooperation.' In a network society, search engines like Google Scholar constitute a nodal point of power, while the mechanism of knowledge production is effectively hidden in the coded mechanisms of the engine, as well as in the unarticulated conventions of scholarly use, such as quality assessment and source presentation. This becomes all the more significant when we look at how other public values of information search, such as relevance and privacy, are incorporated in this new powerful network of knowledge production.

Trusting the shoulders of giants

Electronic library services are a great advance in users' abilities to find and cite sources, but they are never simple mediators between data and knowledge. A few researchers have convincingly argued that electronic databases impact students' and scholars' search behaviour and, vice versa, that interfaces influence the production of knowledge by steering the behaviour of users (Introna and Nissenbaum, 2000; Van Couvering, 2007). One specific effect of online search activity, according to American sociologist James Evans (2008), is that research gets anchored less deeply into past and present scholarship. Paradoxically, as more journal issues are becoming available online, fewer journals and articles are cited. He concludes that online search, even if more efficient, accelerates consensus and narrows the range of findings and ideas built upon. What holds for online resources in general, specifically applies to search engines. Although Google Scholar is obviously far less used than its parent-engine, Google's power over data and information from scholarly sources is amplified by the engine's ability to link up to other layers of data and endless other databases (Hunter et al., 2009). Google's powerful position in the information retrieval market warrants extra vigilance with regards to the reliability, relevance and transparency of its rankings.

Although below I will return to the dangers of having one proprietary engine as a preferred gateway to scientific knowledge construction, I focus first on the mutual shaping of engine technology and user behaviour. As said, the engine lists electronic sources found by its crawlers according to their popularity as a cited source. Popularity in the Google-universe has everything to do with quantity and very little with quality or relevance. To some extent, ranking academic sources through Google Scholar is like ranking the stars: you get what most people voted for, or rather, clicked for. The more a source is linked to, the higher it will end in the ranking of a new query, whatever its value or relevance to a specific research question. Search engine scholarship is thus an ecosystem of reputation where popularity is ranked by an algorithm measuring the relative weight of sources as they are mentioned by random users. To put it simply: queries tend to reward sources already cited over sources that are less well-connected; this 'rich-get-richer' or 'winner-takes all' effect – that is, much-cited sources gain prominence at the expense of sources that are connected to less often – is a well researched yet disputed phenomenon in search engines.[4] The bottom line is that search engines, in crawling references on specific topics, tend to favour groups of highly interlinked sites primarily published by visible (often institutional) sources. This 'chunky' nature of the web is calcified by Google's PageRank technology (Halavais, 2009). Users who continuously trace these same sources by clicking on them, unwittingly amplify the engine's privileging effect. On par with Evans' (2008) findings, several researchers have concluded that the use of online search funnels consensus and narrows the variety of sources and ideas (Bar-Ilan, 2008; Mikki, 2009).

To be sure, there is nothing wrong with 'reputational' ranking of sources per se. As Harvard law professor Cass R. Sunstein (2006) has shown, aggregated information of large numbers of searches generally provides accurate information. And yet, the system is vulnerable to bias and distortion: accuracy and relevance will increase as users are more knowledgeable about the issue at stake, and will decrease when aggregated expert judgements suffer from systematic bias or error (Sunstein, 2006: 41). This general rule may also apply to Google Scholar as a specialized engine which is primarily used by people who share an interest in academic research. Obviously, the engine is intended to be used by experts or scholars, and it could be expected that the advanced expertise of its users renders the search engine more intelligent over time. However, the variety of Google Scholar's users is as wide as the scope of documents it crawls. Just as the engine's ranking does not discriminate between published (peer-reviewed) and non-published scholarly articles, the engine does not distinguish between established scholars and beginning students, or between scientists and lay persons, as users. The engine's motto, 'Standing on the shoulders of giants', should thus be taken with a grain of salt. And while it is acceptable for students to stand on the shoulders of giants to learn the profession, it becomes a different story when Google Scholar functions as a one-stop shop on the road to scholarly knowledge.

For students who fundamentally rely on the search engine for its comprehensive scope of covered sources, it is important to know how bias is already anchored in the selection of databases the engine crawls. As Neuhaus et al. (2006) point out, Google Scholar has an extensive coverage of science and medical databases, open-access databases and single-publisher databases, but it is rather weak in covering social science and humanities databases. Even if American researchers found the proprietary engine to be performing well in specific subject areas in comparison to other public library engines, they still point to the fact that a substantial portion of the citations provided by Google Scholar are incomplete (Walters, 2007; White, 2007). Moreover, data collections are overwhelmingly based in English-language sources – US sources most prominently among them – and thus disproportionately privilege and reinforce the American dominance of search results (Halavais, 2009).

Besides relying on the engine's functionality in terms of selecting sources, users also tend to accept Google's *ranking* at face value. Students with little research experience appear to have remarkable faith in Google's judgement when it comes to the engine's prioritizing of search results. An eye-tracking experiment by a multi-disciplinary team of researchers revealed that college students trust the engine's ability to rank results by their relevance to the query (Pan et al., 2007); they tend to look primarily or even exclusively at the first 10 results displayed on the Google page, and they click on results in higher positions even when the abstracts were deliberately manipulated and were made less relevant to the task. Especially when students are fairly ignorant of the topic they are researching, the potential for misguided trust is exacerbated by the non-egalitarian distribution of information on the result pages:

> Combining users' proclivity to trust ranked results with Google's algorithm, increases the chances that those 'already rich' by virtue of nepotism get filthy rich by virtue of robotic searchers. Smaller, less affluent, alternative sites are doubly punished by ranking algorithms and lethargic searchers. (Pan et al., 2007: 817)

To be sure, Pan's test was executed in the context of Google's general search engine, but the page-popularity mechanism works the same for Google Scholar. Students who trust Google Scholar as an arbiter of indexed and prioritized scientific information are likely to be equally uncritical in assessing and weighing sources' value. Trusting Google's engine as a reliable arbiter of knowledge sources seems to be a widespread default attitude among its users, as Pan et al. conclude. This is, first, because people generally do not regard machines in their technical dimensions as manipulative tools and, second, because they have a blind trust in the people and companies designing and operating those tools. According to Pew-researcher Deborah Fallows (2005: i), 68 percent of all users say that search engines are a fair and unbiased source of information. Even if it is fair to assume that users of Google Scholar are more sophisticated than users of search engines in general, it is doubtful whether they are more knowledgeable about the concept of search engines and how they work.

Users' profound trust in the neutrality of the apparatus resonates with the assumption that search engines like Google Search are not in themselves regulative and they cannot be manipulated. This misconception is most obvious when it comes to search engines in general. While Google Search should be compared to the Yellow Pages, where those included pay for their entries, rather than to telephone directories, most users do not recognize the engine's commercial bias. Fallows (2005: ii) found that 62 percent of all search engine users are unaware of the distinction between results which have been paid for (i.e. they are effectively ads) and those that have not, and only 18 percent can tell the difference between these two types. Manipulation of rankings via paid placement is an inherent part of the Google business (Batelle, 2005) and the ability to distinguish between the two should be an essential part of information literacy training (Nicholson et al., 2006). Indeed, Google *has* to be able to manipulate its ranking mechanisms because the engine faces constant external manipulative actions, such as search optimization and click fraud. The company employs a handful of engineers who are constantly fine-tuning its engine and recalibrating its ranked results. However, it is not manipulation itself that is problematic; it is the lack of transparency in how rankings are controlled which is at odds with public library values. In all fairness, the mechanisms of paid placements and ranking manipulation may be less pertinent to Google Scholar than to its parent engine because Google Scholar is not exactly part of a commercial context. As said before, all Google engines are intricately connected, not only because students tend to switch back and forth between Google Search and Google Scholar, but also through the potential of connecting metadata, which I will elaborate in the next section.

The question of trust hardly applies to a single search engine service, but rather applies to the parent company as a whole. It is not an overstatement to say that, increasingly, we trust a single company to regulate the data needed for the production of scientific knowledge. If that company assumes this enormous responsibility, it should at least give an insight into the basic mechanisms by which its tools execute selection and search, and into the policies that guarantee certain basic values (such as quality and reliability) to its users. From a company that carries 'Do no evil' as its corporate mantra, one may expect more than moral platitudes when it comes to guaranteeing profound public values in the construction of knowledge and science. Yet, according to a recent Gartner report: 'even if all Google-employees make good-faith attempts to follow through on the mandate to do no evil, the ability to execute is not assured' (Hunter et al., 2009: 37). Google's notorious reluctance to disclose information on procedural policies, inclusion of sources, technical specifications of its algorithm and policies on issues concerning security and privacy are a matter for concern. Transparency of procedures and methods for gathering sources are key factors for progress in the area of knowledge production.

Manuel Castells (2009: 46) refers to this tendency to connect various networks of communication as the 'switching' strategy: 'the control of the connecting points between various strategic networks'. Google, as a corporation deploying a large number of technological networks in a variety of subject areas (from geographical mapping and social networking to computational statistics), is a key player in the global distribution of communication power. As a prominent *switcher*, the company handles a large number of specific interface systems that relate and connect a variety of user contexts, which are all crucial to the formation of knowledge. It is through these switching processes, enacted by actor-networks – networks which induce synergizing effects – that knowledge is constructed. Switchers are actors: 'made of networks of actors engaging in dynamic interfaces that are specifically operated in each process of connection' (Castells, 2009: 47). Crucial to the 'unwiring' of network activities is understanding the logic in which these processes are grounded; more specifically, we should examine the local workings of global networks and 'identify the frames in the network that frame the mind' (2009: 431). The next section, therefore, focuses on search engines as *profiling* systems, which, in connection to ranking systems, constitute the heart of automated search and further advance Google's position in the race for knowledge production.

Individual and collective profiling

The 'technological unconsciousness' of most users exceeds their basic unawareness of the limited scope and manipulative ranking mechanisms governing their favourite search engine and their own search behaviour. While search engines pay lip service to some public library values, such as citation analysis and relevance, I have already pointed out how they hardly subscribe to basic values such as neutrality and transparency. In addition, commercial search engines are conspicuously silent when it comes to another fundamental public value in academic knowledge production: policies of privacy and confidentiality protection. These values are extremely relevant when we turn the question 'How do users benefit from Google?' into 'How does Google benefit from its users?'

As mentioned before, search engines, besides their prolific function as ranking tools, are also profiling machines. When talking about profiling machines, we commonly think of social network sites such as Orkut, Facebook or Hyves, but in more than one way, search engines are powerful profiling tools by virtue of the metadata produced – often unwittingly – by their users. Every single search on the Internet leaves traces of its sender: query key words, activity log, date and time, search history, etc. All these data can easily be traced back to personal Internet addresses and thus to specific people. Even if they are not actually used to track web behaviour to specific users, these data can be aggregated to a level that allows for user *profiles*: typical features of people who have shown similar search behaviour.

We are long used to the deployment of aggregated metadata in commercial contexts, for instance through Amazon's automated message: 'Customers who bought this item, also bought….' Metadata are the backbone of all Google's search engines; through the provenance of aggregated profiling techniques, Google is able to collect and connect endless streams of behavioural patterns and reconnect them in ways we can only begin to imagine. Each time any user performs a search through one of its engines (Search, Scholar, Images, Maps, News, YouTube, Orkut, etc.), Google gains information about our individual and collective search behaviour. Its rich metadata collections, and particularly the aggregated profiles resulting from these data, are the company's most valuable asset vis-à-vis advertisers, marketers and, for that matter, any type of agency interested in users' potential interests, whether as consumers, employees or as voters. There are very few legal hurdles to the exploitation of metadata reaped from (free) search, and most users are completely unaware of what the engine's metadata disclose about their identity or behaviour. Some of Google's engines, such as YouTube, offer end user licence agreements (eulas), which explicitly state that users allow the company to peruse metadata for all internal purposes or sell them to third parties in aggregated form (van Dijck and Nieborg, 2009).

It might be argued, at this point, that the perusal of metadata from Google Scholar is not commercially interesting and is therefore irrelevant to the argument laid out in this article. Why would individual search behaviour or even collective profiles of particular types of users (e.g. users in certain geographical areas, universities, subjects, disciplinary fields) be of interest to Google itself or to third parties, whether government or commercial agents? For one thing, public library values require they never disclose the interests of their clients to outsiders and, second, they have to protect every person's right to freedom of information. Search engines have never openly subscribed to these values and are still wrestling with the legal issues of confidentiality and privacy when it comes to the use of metadata. Even if search engine companies argue they have no interest whatsoever in individual metadata records, it is crystal clear that even anonymized logs of individual search data can lead to identifiable people (Halavais, 2009). Recent legal battles with various governments over the right to access search data have led Google Inc. to acquiesce to specific demands of regimes in countries with substantial user markets, such as China. Even less regulated is the protection of metadata against the interests of commercial agencies, for whom information about individual search behaviour may be extremely valuable.

However important this type of personal privacy protection – including the vulnerability of stored metadata to malicious attacks by outsiders and unreliable insiders – it is not my main focus here. More pertinent to my argument is the power of search engines to steer *collective* profiling: collective profiles of users may eventually shape the production of knowledge, even if subtly and non-intentionally. Scientific ideas usually develop through associative thinking, the patterns of which can be traced back to individual scholars, but also to groups of related, interlinked researchers. Search engines, in a way, are global associative memories of information sources: they help trace and store metadata on their usage which may subsequently be analysed to result in user profiles and be connected to other data collections or profiles. For instance, tracking down the search behaviour of (a group of) pharmaceutical researchers and connecting these aggregated data to trend analyses of virus dispersion or flu epidemics, can give insiders of search engine companies a considerable information advantage that may play well on (commercial) stock markets. In fact, by allowing search engine companies to peruse and aggregate their search data, researchers unwittingly give away pieces of data, which, if intelligently combined, may lead to insights they cannot possibly have themselves. Jonathan Poritz (2007: 11) has eloquently explained this process as search engines creating ‘a network structure on the mind-map of the collective unconscious’. He warns of the large unmapped legal and moral territory where issues of what he calls ‘collective privacy’ are unregulated in the virtual bonanza of the Internet. Since collective privacy is even harder to define – and therefore harder to protect – than individual privacy, Poritz calls for vigilance in how search engines, and particularly its market leader, exploit instruments for aggregation and interpretation of metadata to which they have exclusive access.

I am not insinuating that search engine companies in general, or Google in particular, are currently abusing their privileged position in data mining to take advantage of ‘clouds’ of knowledge that individual researchers or groups cannot possibly retrieve. What I am pointing at is a large normative and legal grey area where much is at stake that is, as of yet, unregulated. As Poritz proposes, legislation could be designed to prevent search engine companies and their employees from using aggregated search data for insider trading or by requiring full disclosure of these data in completely anonymized, aggregated form, in order to guarantee a level playing field where private search companies have no unfair advantage over public scholars who actually deliver most of the data. In the current situation, search engine companies have an unfair competitive edge over (public) researchers when it comes to the availability and accessibility of huge data collections, the assessment and interpretation of which is key to the production of knowledge. Or, in the words of Alexander Halavais (2009: 157): ‘In an era in which knowledge is the only bankable commodity, search engines own the exchange floor.’

What is important here is not the identification of concrete social actors who will use (or abuse) their powerful advanced position when it comes to data brokering, but to point out how the potential switching between different networks is a fundamental source of power in a global networked society. Manuel Castells urges in his monumental book *Communication Power* that, in order to understand how power works, we need to ‘look [into] the connections between corporate communication networks, financial networks, cultural industrial networks, technology networks, and political networks’ (2009: 431). If we do not unravel (‘unwire’) how power works in a networked society, we cannot understand, neutralize or counteract dominant forces of knowledge production. In the final section of this article, I want to discuss how the 'technological unconsciousness' of students and young scholars needs to be remedied, not just by information literacy but by an enriched kind of information literacy, which includes insights into the mechanisms and politics of knowledge production.

Enriched information literacy

Now that the use of search engines in academic settings has become part of students' operative mindset, it is even more urgent to raise awareness of what these tools can do and how they work. Most American and European universities teach information literacy as part of (mandatory) courses in basic research skills, educating students to take advantage of the large armamentarium of academic search tools, both electronic and otherwise. Information literacy – generally considered the domain of librarians and library scholars – is defined by the American Library Association as: 'the lifelong ability to recognize the need for, to locate, evaluate and effectively use information'.[5] Depending on the expansiveness of this definition, we need to pause and ask whether our students' ability to evaluate and effectively use information is enough to counteract generally naïve and gullible search habits. In light of the issues discussed above, we clearly should do more than simply teach the mechanics of use if we want to equip students with enough critical stamina to face the ever more intricate technical challenges the information society poses in the 21st century.

A number of librarians and library scholars have already responded to the growing dominance of search engines, particularly Google Search and Google Scholar, by emphasizing the importance of including proprietary engines in their instruction practices and web sites; explaining the differences and limitations of various engines may ensure the effective use of resources by students and inexperienced scholars (Cathcart and Roberts, 2005; Ettinger, 2008; Schmidt, 2007). In keeping with the tradition of teaching information literacy as a particular skill (Grafstein, 2002), the antidote to students' ignorance is teaching them the technicalities of various search methods, including their benefits and drawbacks. Other scholars regard e-literacy to be an integral part of teaching information society, as they expand the definition of information literacy to include socio-technical aspects of search (Tuominen et al., 2005; Wallis, 2005). Teaching the mechanics and technicalities of advanced search is extremely important for students who need to be empowered in their quest for knowledge; the help of librarians in this process is essential, but the daunting task of educating students in information literacy cannot be limited to library and teaching professionals only. Instead, this should be the responsibility of *all* academics concerned with public values related to the production of knowledge. Besides assessing search engines' usefulness and reliability, we should also be able to judge these tools in terms of neutrality, transparency, independence, and their openness when it comes to the right to privacy and confidentiality of their users.

In addition to information literacy, I propose to widen the term to encompass a definition that goes well beyond pedagogical skills and teaching practices, so as to include the economic, political and socio-cultural dimensions of search engines (Crowley, 2005; Williams, 2007). Proprietary search engines substantially shape the road from raw data to scholarly knowledge, while rendering essential processes of weighing, evaluating and contextualizing data into black boxes. To turn information into knowledge, students not only need to be socialized into the various stages of the process, but they should also be enabled to critically analyse the tools that help to construct knowledge. They need to understand how information works on all levels, including the more abstract levels of informational politics (Halavais, 2009; Rogers, 2005). Without a basic understanding of network architecture, the dynamics of network connections and their intersections, it is hard to grasp the social, legal, cultural and economic implications of search engines.

A more complete and profound comprehension of the underlying mechanisms of search may raise students' critical awareness of their own agency. A preference for convenience has never been science's biggest enemy; naivety and indifference are arguably its biggest foes. As Nicholas Carr observes in his book on the world's circuited information: 'We accept greater control in return for greater convenience. The spider's web is made to measure, and we're not unhappy inside it' (2008b: 209). Unawareness of the implications of convenient yet black-boxed tools inevitably leads to more control by owners of search technologies over the production of knowledge. The simpler the interface, the more complex and invisible its underlying algorithms – one only need look at Google's magic square to find proof of this paradoxical message. Raising critical awareness of the underlying mechanisms, even (or perhaps especially) when they are black-boxed, is at the heart of information literacy: our knowledge of the media through which information is channelled and constructed is as indispensable as our knowledge of the human brain. If Google has become the central nervous system in the production of knowledge, we need to know as much as possible about its wiring. Or, as Castells (2009: 431) advises: 'Unwire and rewire. Unwire what you do not get, and rewire what makes sense to you.'

By including social constructivist insights in teaching information literacy, we will equip students to face the increasingly complicated technical systems their information world is made of. We are only at the early stages of a future in which the production of knowledge is increasingly automated, interlinked and defined by (specialized) search engines. As Web 1.0 (read only, one-way traffic) has given way to Web 2.0 (reading and writing, two-way traffic), Web 3.0 will give rise to intelligent systems based on tracking, interpreting and predicting intuitive behaviour of human actors. The development of the semantic web, which invisibly connects a large variety of databases and information systems – from financial data to climate change information – redefines knowledge as the smart connection of numerous layers of automatically generated data streams, a process infinitely more complex than today's search algorithms.[6] While Google is already the nerve centre of our information society, we will soon be facing the emergence of intelligent systems that may well form the heart and mind of the information society. Information networks are the architecture of power distribution, as Castells argues, and search engines are important nodes in the construction and distribution of knowledge. In order to ensure future generations of critical and knowledgeable scholars, we need to teach information literacy enriched with analytical skills *and* critical judgement. The production of scientific knowledge is way too important to leave to companies and intelligent machines.

Notes

[1] See: http://scholar.google.com/intl/en/scholar/about.html (consulted June 2010; emphasis added).

[2] Mikki (2009) compares the free web engine to Web of Science and concludes that Google Scholar performs poorly in terms of advanced information retrieval compared to scholarly databases. Schroeder (2007), comparing the same two tools, gives a comprehensive overview of studies comparing Google Scholar and Web of Science. In a comparative study of Google Scholar, Scopus and Web of Science, Bakkalbasi et al. (2006) conclude that each of these tools has strengths in specific disciplinary searches. A behavioural study comparing Metalib with Google Scholar (Nygren et al., 2006) shows how students' preference for convenience leads to Google Scholar being their first choice.

[3] In 2005, the Online Computer Library Center (OCLC) published a report stating that 89 percent of all college students surveyed begin their information searches with a search engine, versus just 2 percent beginning searches on a university library web site. Markland (2005) finds that over 70 percent of students start with a general search engine, and Google is by far their favourite. Google's popularity dwarfs that of Yahoo, MSNSearch (the two recently announced their cooperation in the search engine Bing!), Gigablast and Ask.com, which lag far behind the market leader. Estimates of what percentage of searches is performed through Google varies from 30 percent to 50 percent (Caldas et al., 2008; Hargittai, 2004, 2007).

[4] For a nuanced discussion of this 'winner-takes-all' effect, see Caldas et al. (2008: 770–2), who argue that the effect increases within a narrower area of search. Their webmetric analyses confirmed the 'cliquishness' of web engine search.

[5] Digital information fluency (DIF) is regarded by the American Library Association as a subset of information literacy, defined as:

the ability to find, evaluate and use digital information effectively, efficiently and ethically. DIF involves knowing how digital information is different from print information; having the skills to use specialized tools for finding digital information; and developing the dispositions needed in the digital information environment (see 21st Century Digital Information Fluency [DIF] project and model, or go to http://wikieducator.org/Digital_information_literacy [consulted June 2010]).

6 As a recent article in the *New York Times* discloses, statisticians are increasingly hired by search engine companies; Google is alleged to employ over 250 statistics specialists who work on connecting various data layers and databases (see Lohr, 2009).

References

Bakkalbasi, Nisa, Kathleen Bauer, Janis Glover and Lei Wang (2006) 'Three Options for Citation Tracking: Google Scholar, Scopus and Web of Science', *Biomedical Digital Libraries* 3(7), URL (consulted August 2009): http://www.bio-diglib.com/content/3/1/7

Bar-Ilan, Judit (2007) 'Google Bombing from a Time Perspective', *Journal of Computer-Mediated Communication* 12: 910–38.

Bar-Ilan, Judit (2008) 'Which h-index? A Comparison of WoS, Scopus and Google Scholar', *Scientometrics* 74(2): 257–71.

Batelle, John (2005) *The Search: How Google and its Rivals Rewrote the Rules of Business and Transformed Our Culture*. New York: Penguin.

Caldas, Alexandre, Ralph Schroeder, Gustavo S. Mesch and William H. Dutton (2008) 'Patterns of Information Search and Access on the World Wide Web: Democratizing Expertise or Creating New Hierarchies?', *Journal of Computer-Mediated Communication* 13: 769–93.

Carr, Nicholas (2008a) 'Is Google Making Us Stupid?', *Atlantic Monthly* July/August, URL (consulted May 2009): www.theatlantic.com/doc/print/200807/google

Carr, Nicholas (2008b) *The Big Switch: Rewiring the World, from Edison to Google*. New York: Norton.

Cathcart, Rachael and Amanda Roberts (2005) 'Evaluating Google Scholar as a Tool for Information Literacy', *Internet Reference Services Quarterly* 10(3): 167–76.

Castells, Manuel (2009) *Communication Power*. Oxford: Oxford University Press.

Caufield, James (2005) 'Where Did Google Get Its Value?', *Libraries and the Academy* 5(4): 555–72.

Crowley, Bill (2005) *Spanning the Theory–Practice Divide in Library and Information Science*. Lanham, MD: Scarecrow Press.

Dijck, José van (2009) 'Users Like You: Theorizing Agency in User-generated Content', *Media, Culture & Society* 31(1): 41–58.

Dijck, José van and David Nieborg (2009) ‘Wikinomics and Its Discontents: A Critical Analysis of Web 2.0 Business Manifestoes’, *New Media & Society* 11(4): 855–74.

Ettinger, David (2008) ‘The Triumph of Expediency: The Impact of Google Scholar on Library Instruction’, *Journal of Library Administration* 46(3): 65–72.

Evans, James A. (2008) ‘Electronic Publication and the Narrowing of Science and Scholarship’, *Science* 321(5887): 395–9.

Fallows, Deborah (2005) ‘Search Engine Users’, Pew Internet and American Life Project’, URL (consulted August 2009): http://www.pewinternet.org/Reports/2005/Search-Engine-Users.aspx

Foucault, Michel (1980) *Power/Knowledge: Selected Interviews and Other Writings 1972–1977*, edited by Colin Gordon. New York: Pantheon.

Gorman, G.E. (2006) ‘Giving Way to Google’, *Online Information Review* 30(2): 97–9.

Grafstein, Ann (2002) ‘A Discipline-based Approach to Information Literacy’, *Journal of Academic Librarianship* 28(4): 197–204.

Halavais, Alexander (2009) *Search Engine Society*. Cambridge: Polity Press.

Hargittai, Eszter (2004) ‘Do You “Google”? Understanding Search Engine Use beyond the Hype’, *First Monday* 9(3), URL (consulted June 2010): http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1127/1047

Hargittai, Eszter (2007) ‘The Social, Political, Economic, and Cultural Dimensions of Search Engines: An Introduction’, *Journal of Computer-Mediated Communication* 12: 769–77.

Hellsten, Iina, Loet Leydesdorff and Paul Wouters (2006) ‘Multiple Presents: How Search Engines Rewrite the Past’, *New Media & Society* 8(6): 901–24.

Hunter, Richard, Richard J. De Lotto, Andrew Frank, Bill Gassmann, Arabella Hallawell, Jay Heiser et al. (2009) *What Does Google Know?* Report G00158124. Stamford, CT: Gartner Research.

Introna, Lucas D. and Helen Nissenbaum (2000) ‘Shaping the Web: Why the Politics of Search Engines Matters’, *The Information Society* 16: 169–85.

Jacso, P. (2008) ‘Google Scholar Revisited’, *Online Information Review* 32(1): 102–14.

Latour, Bruno (2005) *Reassembling the Social: An Introduction to Actor-Network Theory.* Oxford: Oxford University Press.

Lohr, Steve (2009) ‘For Today’s Graduate, Just One Word: Statistics’, 5 August, URL (consulted June 2010): http://www.nytimes.com/2009/08/06/technology/06stats.html

Markland, Margaret (2005) ‘Does the Student’s Love of the Search Engine Mean that High-quality Online Academic Resources Are Being Missed?’, *Performance Measurement and Metrics: The International Journal for Library and Information Services* 6(1): 19–31.

Mikki, Susanne (2009) ‘Google Scholar Compared to Web of Science: A Literature Review’, *Nordic Journal of Information Literacy in Higher Education* 1(1): 41–51.

Neuhaus, Chris, Ellen Neuhaus, Alan Asher and Clint Wrede (2006) ‘The Depth and Breadth of Google Scholar: An Empirical Study’, *Portal: Libraries and the Academy* 6(2): 127–41.

Nicholson, Scott, Tito Sierra, U. Yeliz Eseryel, Ji-Hong Park, Philip Barkow, Erika J. Pozo et al. (2006) ‘How Much of It Is Real? Analysis of Paid Placement in Web Search Engine Results’, *Journal of the American Society for Information Science* 57(4): 448–61.

Nygren, Else, Glenn Haya and Wilhelm Widmark (2006) ‘Students’ Experience of Metalib and Google Scholar’, *Online Information Review* 31(3): 365–75.

Online Computer Library Center (2005) *College Students’ Perceptions of Libraries and Information Resources*. Dublin, OH: OCLC, URL (consulted August 2009): http://www.oclc.org/reports/perceptionscollege.htm

Pan, Bin, Helene Hembrooke, Thorsten Joachims, Lori Lorigo, Geri Gay and Laura Granka (2007) ‘In Google We Trust: Users’ Decisions on Rank, Position, and Relevance’, *Journal of Computer-Mediated Communication* 12: 801–23.

Poritz, Jonathan (2007) ‘Who Searches the Searchers? Community Privacy in the Age of Monolithic Search Engines’, *The Information Society* 23(5): 383–9.

Rogers, Richard (2005) *Information Politics on the Web*. Cambridge, MA: MIT Press.

Schmidt, Janine (2007) ‘Promoting Library Services in a Google World’, *Library Management* 28(6): 337–46.

Schroeder, Robert (2007) ‘Pointing Users Toward Citation Searching: Using Google Scholar and Web of Science’, *Portal: Libraries and the Academy* 7(2): 243–8.

Sunstein, Cass R. (2006) *Infotopia: How Many Minds Produce Knowledge*. Oxford: Oxford University Press.

Taylor, Stephanie (2007) ‘Google Scholar – Friend or Foe?’, *Interlending & Document Supply* 35(1): 4–6.

Tuominen, Kimmo, Reijo Savolainen and Sanna Talja (2005) ‘Information Literacy as a Sociotechnical Practice’, *Library Quarterly* 75(3): 329–45.

Van Couvering, Elizabeth (2007) ‘Is Relevance Relevant? Market, Science, and War: Discourse of Search Engine Quality’, *Journal of Computer-Mediated Communication* 12, URL (consulted June 2010): http://jcmc.indiana.edu/vol12/issue3/vancouvering.html

Wallis, Jake (2005) ‘Cyberspace, Information Literacy and the Information Society’, *Library Review* 54(4): 218–22.

Walters, William H. (2007) ‘Google Scholar Coverage of a Multidisciplinary Field’, *Information Processing and Management* 43(4): 1121–32.

White, Bruce (2007) ‘Examining the Claims of Google Scholar as a Serious Information Source’, *New Zealand Library & Information Management Journal* 50(1): 11–24.

Williams, Genevieve (2007) ‘Unclear on the Context: Refocusing on Information Literacy’s Evaluative Component in the Age of Google’, *Library Philosophy and Practice* 6, URL (consulted August 2009): http://www.encyclopedia.com/Library+Philosophy+and+Practice/publications.aspx?date=200706&pageNumber=1

JOSÉ VAN DIJCK is a Professor of Media and Culture at the University of Amsterdam where she is currently the Dean of Humanities. Her research areas include media and science, (digital) media technologies, popularization of science and medicine, and television and culture. She is the author of several books, including *Manufacturing Babies and Public Consent: Debating the New Reproductive Technologies* (New York University Press, 1995), *ImagEnation: Popular Images of Genetics* (New York University Press, 1998) and *The Transparent Body: A Cultural Analysis of Medical Imaging* (University of Washington Press, 2005). Her latest book, *Mediated Memories in the Digital Age,* in which she theorizes the relationship between media technologies and cultural memory, was published by Stanford University Press (2007). *Address*: University of Amsterdam, Faculty of Humanities, Spuistraat 210, 1012 XT Amsterdam, The Netherlands. [email: j.van.dijck@uva.nl]