

Google, Public Libraries, and the Deep Web

Abstract: A February 2012 Pew Internet survey found that 91% of online adults use search engines to find information on the web (Purcell, Brenner, & Rainie, 2012). When asked which search engine they use most often, 83% of search users say Google.

This paper outlines a challenge to public libraries, which have a mandate to facilitate public access to information. A preliminary study reveals that many public libraries in Canada currently block their catalogues from Google's indexing robots, effectively rendering their resources invisible to the majority of Internet searchers. This fact undermines the purpose and functionality of public libraries' Online Public Access Catalogues (OPACs). This paper will address the problem of dynamic content for search engine robots, and suggest solutions, as well as illustrating the need for future studies.

About the Author: Alieda is a MLIS candidate from Victoria, British Columbia. She is the Digital Publications Officer for the School of Information Management Student Association and an Editorial Chair for the Dalhousie Journal of Interdisciplinary Management. She has also served as the Fundraising Chair for Librarians Without Borders and is a founding member of FYI Halifax, an action-oriented community group interested in local literacy projects and information activism. Her professional interests lie in public libraries, particularly collection development, community outreach, and emerging technologies.

Introduction

A February 2012 Pew Internet survey found that 91% of online adults use search engines to find information on the web (Purcell, Brenner, & Rainie, 2012). When asked which search engine they use most often, 83% of search users say Google.

These numbers illustrate a grave weakness for many public libraries, whose Open Public Access Catalogues (OPACs) are often hidden in the so-called “deep” Web, out of reach of Google’s indexing robots. Public library resources that are not indexed by search engines cannot populate on search engine results pages (SERPs), so a searcher would need to begin on a library homepage in order to access these resources. Unfortunately, while 91% of online adults use search engines to find information online (Purcell, Brenner, & Rainie, 2012), less than 1% begin their search on a library homepage (OCLC, 2010). This means that publically accessible resources from many public libraries are effectively invisible to the majority of Internet searchers.

Unfortunately, this problem is widespread. Many public library OPACs in Canada are not indexed by search engines; in fact, several actively block search engines from accessing and indexing their catalogues. The invisibility of public library resources on SERPs impairs public access to information, while potentially contributing to libraries’ perceived irrelevance in the digital age. Surprisingly, this problem has not been explored often in library and information science literature.

This paper will explore the issue of search engine visibility, including the importance of Google, the existence of the deep Web, and the problem of surfacing public library resources. It will address the problem of dynamic content for search engine robots, and suggest solutions, such as the creation of stable URLs and sitemaps. Finally, this paper will illustrate the need for future research.

The Importance of Google

There is no doubt that Google has made its mark on society. *Newsweek* reporter Steven Levy (2011) writes that Google has singlehandedly “changed the way people accessed information, changed the way they *thought* about information” (p. 2), and calls the company “one of the

most important contributors to the revolution of computers and technology that marked a turning point in civilization” (p. 3). Indeed, as Ken Auletta (2009), author of *Googled: The end of the world as we know it*, points out: today we don’t search for information – we Google it.

According to Blankenship (2009), Google has been synonymous in the public mind with Internet searching since at least 2002. Cahill (2009) goes one step further, pointing out that for many people, Google is synonymous with not just Web searching, but with research itself. She writes that many people “now believe that finding the answer to a question is as simple as going to a single webpage, typing in a few keywords, and seeing what comes back” (Cahill, 2009, p. 68).

Google’s multi-algorithmic approach to Internet searching, “which combines full text, metadata, and popularity measures,” is “the most successful approach any search engine has taken to producing relevant results” (Cahill, 2009, p. 68). In fact, Google’s algorithms have been so successful that today, most Internet searchers do not look past the first page of results (iProspect, 2008; Jansen & Spink, 2006). Given the central role that search engines play in information search and retrieval online, it is no surprise that the importance of maintaining high visibility on SERPs has steadily increased over time (Onaifo & Rasmussen, 2013, p. 87-88; iProspect, 2008, p. 6). Today, companies and web organizations of all types covet a place in the highly-visible first few results in a given Google search.

Search Engine Optimization

PageRank (named for Google co-founder Larry Page) lies at the heart of Google’s success as a search engine. PageRank is “based on a concept familiar to every library school student: citation analysis” (Cahill & Chalut, 2009, p. 237). In much the same way that scholarly impact is measured by citation counts, Google ranks its search results based on the number of times each page is linked to from other pages. However, this simple vetting process is actually more complex, because “PageRank not only measures the number of times an individual page is linked to, but also weights the relative importance of the pages that are linking to it” (Cahill & Chalut, 2009, p. 237).

As soon as PageRank was released, “webmasters and website owners set to figuring out how to make it work to their advantage” (Cahill & Chalut, 2009, p. 237). Search engine optimization

(SEO), the practice of manipulating the system in order to create a webpage “that will appear as high as possible on the search engine results page (SERP) for a given keyword search,” is currently “one of the key aspects of successfully marketing products and services on the modern web” (Cahill & Chalut, 2009, p. 235).

While SEO is used by nearly every web organization, there are also numerous “Black Hat” SEO techniques used by illicit entities in order to beat the PageRank system. Keyword stuffing, site scraping, doorway pages, link farms, and comment spam are some examples of such techniques (see Cahill & Chalut, 2009). These methods are used “to lure searchers into spam sites under false pretenses” (Cahill & Chalut, 2009, p. 239), or to “farm” backlinks to a given site, raising its SERP rank. In other words, despite Google’s pervasiveness, “it still has many of the search engine’s inherent frailties when it comes to sifting relevant material out of the masses” (Cahill, 2009, p. 68).

However, the information examined thus far has ignored the largest problem for search engines like Google: the existence of the deep, or hidden, Web. Unbeknownst to most Web users, the majority of Web content is actually not indexed by search engines at all, and exists, “for a variety of technical, political, and economic reasons” (Devine & Egger-Sider, 2005, p. 91), out of the reach of Google’s Web crawlers. Since these documents will never populate in a Google search, they are virtually impossible to find unless you know exactly where to look.

What is the Deep Web?

Bergman (2001) writes that searching on the Internet today can be compared to “dragging a net across the surface of the ocean” (p. 1; see Figure 1). The “surface” or “visible” Web contains all of the content that most users would consider “the Internet,” and yet the deep Web is many orders of magnitude larger. To put the scale of the deep Web in perspective, it contains 7,500 terabytes of information, compared to the 19 terabytes of information on the surface Web (Bergman, 2001). Put another way, the hidden Web contains 550 billion documents, compared to the estimated 2.5 billion documents currently indexed by all search engines (and Google itself has indexed just 1.5 billion of these) (Bergman, 2001). In other words, although most Web users are only aware of the content presented to them via Google,

the deep Web is estimated to be at least 500 times larger than the commonly defined World Wide Web (Bergman, 2001).

The visible Web is made of pages that can be indexed by search engine crawlers, also known as “robots” or “spiders.” Google’s web crawlers access webpages via links from other pages. This means that if a webpage “has no links pointing to it from any other page on the Web, a search engine crawler can’t find it” (Sherman & Price, 2003, p. 284). These “disconnected” pages form the most basic part of the deep Web.

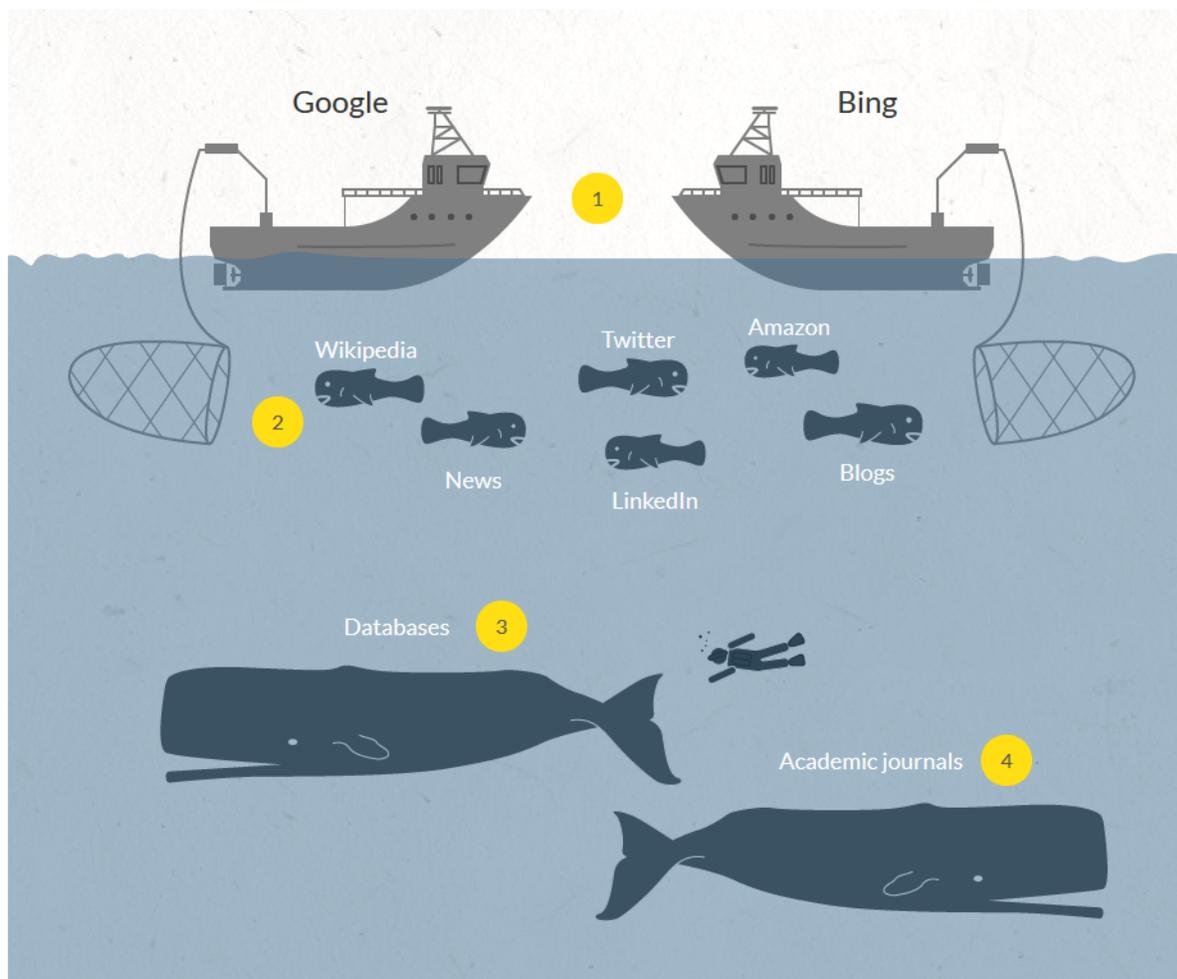


Figure 1: The deep Web. Retrieved from: <http://money.cnn.com/infographic/technology/what-is-the-deep-web/>

Many scholars call these masses of “disconnected” Web pages the “opaque” Web, since this portion of the deep Web could be indexed, but is not (Sherman & Price, 2003, p. 291). If

desired, the author of a disconnected Web page could simply use a search engine's "Submit URL" feature to request that a crawler visit and index the page (Sherman & Price, 2003, p. 295).

As Google's robots crawl the web, the first thing they attempt to determine is "whether access to pages on a server it is attempting to crawl is restricted" (Sherman & Price, 2003, p. 288). Blocked pages make up a significant portion of the deep Web, and scholars often call these kinds of websites the "private" Web, as they have been "deliberately excluded from inclusion in search engines" by their authors (Sherman & Price, 2003, p. 296). Password-protected pages (for instance, private e-mail accounts) make up a portion of the private Web.

Websites may also be hidden in the private Web due to the inclusion of Robots Exclusion Protocol (robot.txt) files or "noindex" metatags. The Robots Exclusion Protocol is:

a set of rules that enable a Webmaster to specify which parts of a server are open to search engine crawlers, and which parts are off-limits. The Webmaster simply creates a list of files or directories that should not be crawled or indexed and saves this list on the server in a file named robots.txt. This optional file, stored by convention at the top level of a Web site, is nothing more than a polite request to the crawler to keep out, but most major search engines respect the protocol and will not index files specified in robots.txt (Sherman & Price, 2003, p. 288).

Since most indexing robots honor robots.txt files that disallow search engine crawlers from indexing specified pages, this protocol is sometimes used by website authors to disable the harvesting of a website (or particular directories of a website) by search engines (Boston, 2005, p. 6). Additionally, directives embedded in HTML pages such as "noindex" and "follow" can be used to control harvesting (Boston, 2005, p. 6).

Aside from the private Web, "[t]he biggest technical hurdle search engines face lies in accessing information stored in databases" (Sherman & Price, 2003, p. 286). Sherman and Price (2003) call this a "huge problem," because there are thousands – and perhaps millions – of content-rich documents from "universities, libraries, associations, businesses, and government agencies" (p. 286), among other scientific and cultural institutions, hidden in this portion of the deep Web. The existence of the "truly Invisible Web" (Sherman & Price, 2003, p.

297) has been called “the most interesting data integration problem imaginable” by Alon Halevy, a former computer science professor at the University of Washington who is now leading a team at Google that is trying to solve the deep Web conundrum (Wright, 2009).

Traditionally, even publically available information that is stored in relational databases cannot be indexed by Web crawlers, even if those databases are hosted by Web interfaces which are, themselves, indexed. That is because these interfaces include search forms, which require input from a user in order to generate a unique page of results based on a searcher’s query. Search engine robots are not built to generate this kind of dynamic content. Sherman and Price (2003) elaborate:

Although these interfaces provide powerful tools for a human searcher, they act as roadblocks for a search engine spider. Essentially, when an indexing spider comes across a database, it’s as if it has run smack into the entrance of a massive library with securely bolted doors. A crawler can locate and index the library’s address, but because the crawler cannot penetrate the gateway it can’t tell you anything about the books, magazines, or other documents it contains. (p. 286).

Exposing Library OPACs to the Surface Web

Library OPACs are often characterized by dynamic content, which means that a unique page of results is generated each time a user interacts with the catalogue search form. Since Google’s crawlers are not designed to interact with databases in this way, the existence of dynamic content can explain, to some extent, why public library catalogues have traditionally been hidden in the deep Web.

On the other hand, today it is common for each record in a library’s catalogue to be given its own static, persistent URL (see, for example, Halifax Public Libraries, 2014). The main reason why these individual URLs are not indexed by search engine robots is either because they are “disconnected” from other pages, with no links from the surface Web, or because the library website administrator has blocked search engine crawlers from the catalogue records using the Robot Exclusion Protocol.

To solve the problem of disconnected pages in order to make individual library catalogue records accessible to search engine robots, Kumar (2012) recommends the use of a sitemap: essentially, “a plain-text list of all the URLs from a particular website” (p. 495).

In November 2002, the National Library of Australia (NLA) began exposing its digital collections to search engines using sitemaps (Boston, 2005). Each item in the library’s digital collections had been given a “persistent identifier” and a “stable web address,” so exposing these resources to search engines simply involved the creation of “a series of web pages, each containing 100 links,” that resolved to the collection items (Boston, 2005, p. 3). These sitemap pages used the robot directive “noindex, follow” in order to direct search engine harvesters to follow the links to content, but not to index the list pages themselves. Finally, the lists of URLs were connected to the library’s infrastructure page in order to “provide a starting point for search engine harvesters” (Boston, 2005, p. 3). Alternately, these URL lists could have been submitted to search engines in order to request that a crawler visit and index the list pages (see Kumar, 2012; Sherman & Price, 2005).

It is important to note that, after exposing their digital collections to search engine harvesters, the NLA witnessed a “marked increase” in their use: over the next two years, from July 2002 to May 2004, the growth in the usage of their digital collections was 370%, “almost three times the 130% experienced by the NLA website in the same period” (Boston, 2005, p. 4). Notably, “a marked increase in referrals from Google” was seen, from 300 to 4,000 referrals per month in December 2002, rising to 67,000 referrals per month by August 2003 (Boston, 2005, p. 4).

Marshall Breeding (2014) agrees that, in order to expose database content to the surface web, each resource in a library catalogue “should be associated with a stable and simple URL” (p. 25). He also notes that some content management systems “can automatically generate a sitemap.xml document,” which will provide “a systematic listing of all the unique pages available” (p. 25); the name and location of this file, he writes, should be placed in the robots.txt document, then manually submitted to Google.

Aside from these examples, however, there is a surprising dearth of literature in library and information science about the problem of search engine visibility for libraries, and only a handful of studies that explore libraries’ application of SEO exist (Onaifo & Rasmussen, 2012).

This must change. The purpose of a public library OPAC is to provide online access to public information from the library. The fact that so many OPACs are unnecessarily hidden in the deep Web undermines that purpose.

It can sometimes appear that librarians are looking in the wrong direction. For instance, Devine & Egger-Sider (2004) discuss ways in which reference librarians can leverage their knowledge of the deep Web in order to impress patrons and “promote interest in library [reference] services” (p. 265). They point out that reference librarians can pique patrons’ interest by making reference to the hidden Web; they recommend educating library users about Google’s imperfections before dazzling them with research expertise. The authors do not seem to consider the weakness with this approach: it relies on impressing one user at a time as they physically approach the reference desk. And while the authors express awareness of the fact that their own library’s resources are hidden in the deep Web (Devine & Egger-Sider, 2004, p. 266), they do not appear to consider the idea that “interest in library services” could be better promoted by exposing these resources to Internet search engines, thus increasing information access and collection use.

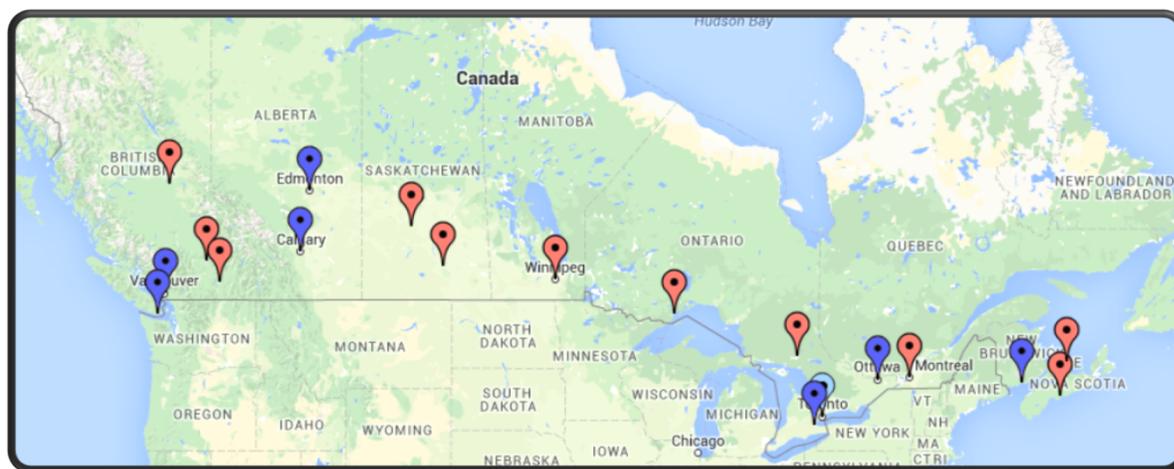
While interest in exposing public library OPACs to the surface Web appears to be minimal, a small hard core of information professionals has been calling for the inclusion of library catalogues in the surface Web for at least a decade. Boston (2005) argues that “[e]xposing the deep web is highly relevant to libraries which traditionally attempt to make their collections as easily accessible to as wide a range of users as possible” (p. 2). He also points out that “there is a growing expectation, especially among younger Internet users, that the high quality collections held by libraries and other cultural institutions should be accessible through Internet search engines” (Boston, 2005, p. 2). Ludwig (2003) agrees, writing:

Librarians must bring the deep, invisible web to the surface so that our public can discover our resources through a query on a popular search engine. While using proprietary database technology to automate content generation and manage lists is efficient, it ultimately undermines the openness and ease of access that once made the Web great. [... the public] would be better served by library content on persistent pages that continue to exist beyond the end of a database search. This would help them find information where they look for it – online. (p. 8).

It would seem, however, that information professionals in the past decade have been slow to take up this call.

Canadian Library Catalogues & the Deep Web

A cursory look through the robots.txt files of Canada's major public libraries will show that many of them, including the public libraries of Calgary, Regina, Saskatoon, Winnipeg, and Halifax, have blocked search engine crawlers from indexing their catalogues (pictured below Figure 2).



 Indexed

 Not indexed

Figure 2: Search engine indexing of Canada's major public libraries' catalogues

As one example, Winnipeg Public Library's robots.txt file (Figure 3) blocks all search engine robots from entering its catalogue. (The file can be viewed at <https://winca.ent.sirsiidynix.net/robots.txt>.)

```
User-agent: *  
Disallow: /
```

Figure 3: Winnipeg Public Libraries' robot.txt file

The “User-agent: *” command means that this section applies to all robots, and the “Disallow: /” tells the robot that it should not visit any pages on the site (Robots.txt, 2007).

Library catalogues may not be indexed by search engines for several reasons. First, a robots.txt file may block search engine crawlers from indexing the catalogue (as above); second, the catalogue may consist of dynamic content, which presents a technical, although not an insurmountable, challenge; and finally, if persistent URLs for each catalogue record exist, they are likely disconnected from other documents on the surface Web.

The creation of a stable, permanent URL for each catalogue item is the first step to exposing a library OPAC to the surface web. A sitemap, detailing the links to a library’s catalogue items, would surface these individual URLs. Concerns about taxing the server with page requests from search engine robots could also be addressed by specifying a rate at which page requests could be made, and instilling a crawl delay.

For instance, BiblioCommons (to the best of my knowledge, the only commercially available public library catalogue platform that allows search engine indexing) specifies a crawl delay of 17 seconds and limits requests to a rate of one per 500 milliseconds (Figure 4).

BiblioCommons also disallows search engine crawlers from indexing certain pages and specifies a window of time for indexing (in the early morning, when the public library’s catalogue has low human traffic) (Figure 4). It should also be noted that BiblioCommons does not use a sitemap, opting instead to allow search engine crawlers to follow the link for each item in the catalogue in order to generate a page for each (Artlitsch, 2013).

```
User-agent: *
Crawl-delay: 17
Request-rate: 1/500ms
Disallow: /layouts/
Disallow: /holds/select_hold/
Disallow: /item/report_match/
Disallow: /list/share/
Disallow: /info/select_library/
Disallow: /monitor/
Disallow: /item/digital_availability/
Allow: /*
Visit-time: 0200-0800
```

Figure 4: BiblioCommons robot.txt file

Because their catalogue items are indexed, resources from libraries using the BiblioCommons platform often populate among the top search results for a query involving the library's home city and an item in their collection (Figure 5).

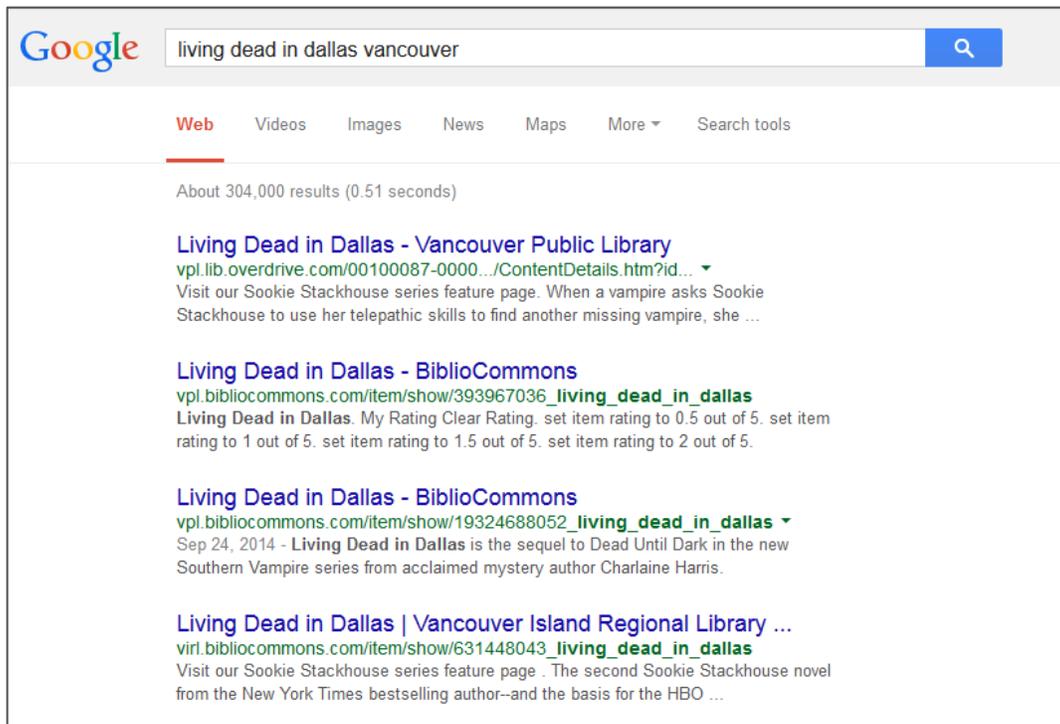


Figure 5: Book title & city name search results from Vancouver Public Library (BiblioCommons)

Google as the Competition

The problem of surfacing the deep Web is non-trivial to those who have a stake in public information access – especially given Internet users' clear preference for using Google. While there are technical hurdles involved in exposing public libraries' catalogues to the surface Web, these challenges are by no means insurmountable.

However, it is possible that some information professionals may be somewhat *willfully* disinterested in this issue due to a commonly held view that libraries “are in competition with other entities that provide information access to students, scholars, and the general public” (Anderson, 2005, p. 29). In other words, far from exposing their catalogues to search engines or employing SEO techniques to ensure that their resources vie for a favourable position on

SERPs, librarians tend to see Google as a rival and a threat (see Bell, 2004; Blankenship, 2009; Devine & Egger-Sider, 2004; Ettinger, 2008; Giglierano, 2008).

This anti-Google attitude is especially apparent today in academic librarians' wariness of Google Scholar. While this service will "easily provide a small but representative selection of scholarly research on most subjects" (O'Leary, 2005, p. 35), ranked by citation count and relevance, and often available in full-text, Google Scholar's "sheer ease of use" and "interdisciplinary reach" (Ettinger, 2008, p. 67) has led many academic librarians to worry that it might "supplant the expensive subscription databases purchased with libraries' ever-shrinking budgets" (Giglierano, 2008, p. 101).

In a 2007 survey of OhioLINK university libraries, for instance, less than half (32%) provided a link to Google Scholar on their websites (Giglierano, 2008, p. 103). The most frequent reason given for not doing so was "wanting students to use the databases the library had paid for" (Giglierano, 2008, p. 104). The general fear seems to be that if the university library provided a link to Google Scholar, students might not use anything else. Cothran (2011) points out that "concerns linger that Google Scholar is drawing students away from libraries to outside resources" (p. 293), and, due to its imperfections, it has been called "a tool of last resort" (Badke, 2013). In other words, "[w]hether they are prepared to acknowledge it or not, information professionals appear threatened by Google Scholar's capacity to upset the traditional model of library service" (Ettinger, 2008, p. 68).

While academic librarians worry about Google Scholar, and reference librarians are feeling threatened in a world where anyone, anywhere, can find answers to reference questions using Google, librarians of all backgrounds are also concerned about their future relevance in a world where Google Books is undertaking a large-scale project to make every book ever published available in digitized form (Doughetery, 2010).

To some, it is worrying that a private corporation has so quickly gained control (as both the harvester and the gatekeeper) of the world's information – arguably the most precious commodity of the digital age. Additionally, research shows that search engines' algorithms return incomplete and biased results. Due to personalization features, for example, "results in

a search engine differ per user,” and this has the potential to create a “monoculture,” in which users get trapped in their individual “filter bubbles” (Bozdag, 2013, p. 209).

The danger of Google’s pervasiveness as an information resource lies in the fact that, since most searchers do not look past the first page of search results (iProspect, 2008), the company has gained considerable influence over which information is consumed by the public. Many people “rely on Google’s results lists to tell them which site is preferred” (Cahill, 2009, p. 68), but this confidence may be misplaced, considering that Google has an economic imperative to push certain results and not others. (This can be seen in the fact that MapQuest results do not populate in Google searches for driving directions: the company wants its searchers to use Google Maps.)

It is worth noting that Google is more than just a search engine. Its subsidiary YouTube accounts for two-thirds of all Web video traffic (Auletta, 2009, p. xi), and the company’s advertising revenues account for 49% of the digital ad market worldwide (compared to Facebook’s 7%) (eMarketer, 2013). Google is involved in satellite imagery, mobile phones, e-mail, energy generation, cloud storage, web-based and mobile applications, and more (Levy, 2011).

These numbers expose Google as an innovative, far-reaching, and highly profitable company; but, as a corporation, it is motivated by profit, and not by a mandate to provide free access to information. Its mission, “to organize the world’s knowledge and make it universally accessible and useful” (Google, n.d.), reflects the company’s goal to provide the best search results to its customers. This goal supports, rather than negates, the economic interests at stake.

With this in mind, it is perhaps unsurprising that some librarians today are sounding an alarm. Quite simply, there are those in the field who fear “that Google will make libraries obsolete” (Miller, 2005, p. 2). On the other hand, libraries’ competition is not just Google itself, but Ask.com, Wikipedia, How Stuff Works, WebMD, and other commercial websites at the top of search results whose information answers the needs of billions of queries a day. Public libraries must move beyond criticism of Google and begin to influence the information environment online in order to better serve the information needs of the public. The most

straightforward way to secure “a place at the table” is to expose library resources to the surface Web and participate in search engine optimization.

Conclusion

If I type the word “pizza” into Google’s search bar, my results page populates with seven local pizza places, complete with addresses, phone numbers, reviews, and mapped locations. In other words, Google “knows” where I live, and it “predicts” that I am interested in eating pizza in my home city. It gives me all the information I need to pursue this goal.

Given the intuitive nature of Google’s search results, it seems inexpressibly unfortunate that, when I search for the word “genealogy,” for example, my results include Ancestry.com, but do not include any reference to the fact that my local library offers free access to this fee-based service to patrons. Similarly, if I was looking for information on gardening, baking, making crafts, playing Bridge, or loading library books onto my e-reader, wouldn’t it be nice if the local public library’s many programs and other resources about these topics populated in my search results?

Allowing search engines to index library catalogues and employing SEO techniques are strategies that could improve public access to library resources. Additionally, exposing library catalogues to the surface Web is the first step in showcasing the library’s high-quality resources where the public is looking – on Google. Further SEO could increase collection use, boost program attendance, and facilitate the use of library services; it could also increase the library’s relevance in the eyes of the public.

Considering that 91% of adults use Internet search engines to find information (Purcell et al, 2012), while less than 1% of them begin their search on a public library website (OCLC, 2010), the fact that so many public libraries in Canada block their catalogues from search engines is astonishing, especially for an institution which has a mandate to facilitate public access to information. Future research should investigate the reason behind public libraries’ decisions to block search engine crawlers from their collections. Additionally, there is a need for empirical studies that measure the cost and impact of catalogue exposure for public libraries.

References

- Anderson, R. (2005). The (uncertain) future of libraries in a Google world: Sounding an alarm. Miller, W. & Pellen, R. M. (Eds.). *Libraries and Google*. (pp. 23-29). Binghamton, New York: The Haworth Information Press.
- Artlitsch, K., & O'Brien, P., and Rossmann, B. (2013). Managing search engine optimization: An introduction for library administrators. *Journal of Library Administration*, 53(2-3), 177-188. doi:10.1080/01930826.2013.853499
- Auletta, K. (2009). *Googled: The end of the world as we know it*. New York: Penguin Group.
- Badke, W. B. (2013). Coming back to Google Scholar. *Online Searcher*, 37(5), 65-67.
- Bell, S. (2004). The infodiet: How libraries can offer an appetizing alternative to Google. *Chronicle of Higher Education*, 50(24), B15.
- Bergman, M. K. (2001). White paper: The deep web: surfacing hidden value. *Journal of Electronic Publishing*, 7(1), 1-37.
- Blankenship, E. F. (2009). Who holds the keys to the web for libraries? Miller, W. & Pellen, R. M. (Eds.). *Googlization of Libraries*. (pp. 54-65). New York: Routledge.
- Boston, T. (2005). Exposing the deep web to increase access to library collections. Proceedings of the *11th Australian World Wide Web Conference*, Gold Coast, Australia.
- Bozdag, E. (2013). Bias in algorithmic filtering and personalization. *Ethics and Information Technology*, 15, 209-227. doi: 10.1007/s10676-013-9321-6
- Breeding, Marshall. (October 2014). Enhancing discovery through better exposure. *Computers in Libraries*, 34(8), 24-26.
- Cahill, K. (2009). An opportunity, not a crisis: How Google is changing the individual and the information profession. In Miller, W. & Pellen, R. M. (Eds.). *Googlization of Libraries*. (pp. 66-74). New York: Routledge.
- Cahill, K. & Chalut, R. (2009). Optimal results: What libraries need to know about Google and search engine optimization. *The Reference Librarian*, 50(3), 234-247. doi: 10.1080/02763870902961969
- Cothran, T. (2011). Google Scholar acceptance and use among graduate students: A quantitative study. *Library & Information Science Research (07408188)*, 33(4), 293-301. doi: 10.1016/j.lisr.2011.02.001

- Devine, J., & Egger-Sider, F. (2004). Beyond Google: The invisible Web in the academic library. *The Journal of Academic Librarianship*, 30(4), 265-269. doi: 10.1016/j.acalib.2004.04.010
- Devine, J., & Egger-Sider, F. (2005). Google, the invisible Web, and librarians: Slaying the research Goliath. *Internet Reference Services Quarterly*, 10(3-4), 89-101. doi: 10.1300/J136v10n03_09
- Dougherty, W. (2010). The Google Books project: Will it make libraries obsolete? *The Journal of Academic Librarianship*, 36(1), 86-89. doi: 10.1016/j.acalib.2009.12.002
- Ettinger, D. (2008). The triumph of expediency: The impact of Google Scholar on library instruction. *Journal of Library Administration*, 46(3-4), 65-72. doi: 10.1300/J111v46n03_06
- eMarketer (2013). Google takes home half of worldwide mobile Internet ad revenues. Retrieved from: <http://www.emarketer.com/Article/Google-Takes-Home-Half-of-Worldwide-Mobile-Internet-Ad-Revenues/1009966#sthash.7qKaKUfx.dpuf>
- Giglierano, J. (2008). Attitudes of OhioLINK librarians toward Google Scholar. *Journal of Library Administration*, 47(1-2), 101-113. doi: 10.1080/01930820802110951
- Google. (n.d.). Google's Mission. Retrieved from: <https://www.google.com/about/company/>
- Google. (2012). Code of Conduct. Retrieved from: <http://investor.google.com/corporate/code-of-conduct.html>
- Halifax Public Libraries (2014). Pride and Prejudice. Retrieved from: <http://discover.halifaxpubliclibraries.ca/?itemid=|library/m/halifax-horizon|200183>
- iProspect (2008). Blended search results study. Retrieved from: http://api.ning.com/files/uHeWp4sJV-TxTYvYOHkwT1pUVuDa91h0kYXJ9yA73NvAkSVbBraKI8y8ey7VxGYVIHF5HTiLz8CGwzA7nj3qlsXShUYJzht5/researchstudy_apr2008_blendedsearchresults.pdf
- Jansen, B.J. & Spink, A. (2006). How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing and Management*, 42, 248-263. doi: 10.1016/j.ipm.2004.10.007
- Kumar, V. (2012). Exposing library catalogues to search engines. *Journal of Library & Information Technology*, 32(6), 493-498. doi: <http://dx.doi.org/10.14429/djlit.32.6.284>
- Levy, S. (2011). *In the Plex: How Google thinks, works, and shapes our lives*. New York: Simon & Schuster.

- Ludwig, M. (Winter 2003). Breaking through the invisible web. *NetConnect*, 8-10.
- Miller, W. (2005). Introduction: Libraries and their interrelationships with Google. Miller, W. & Pellen, R. M. (Eds.). *Libraries and Google*. (pp. 1-4). Binghamton, New York: The Haworth Information Press.
- OCLC (2010). Perceptions of libraries, 2010: Context and community. Retrieved from: http://www.oclc.org/content/dam/oclc/reports/2010perceptions/2010perceptions_all.pdf
- O'Leary, M (2005). Google Scholar: What's in it for you? *Information Today*, 22(7), 35-39.
- Onaifo, D. & Rasmussen, D. (2012). Increasing libraries' content findability of the web with search engine optimization. *Library Hi Tech*, 31(1), 87-108. doi: 10.1108/07378831311303958
- Pagliery, J., & Yellin, T. (2014, March 10). What is the Deep Web? [Infographic]. *CNN Money*. Retrieved from <http://money.cnn.com/infographic/technology/what-is-the-deep-web/>
- Purcell, K., Brenner J., & Rainie, L. (2012). Search engine use 2012. Pew Research Internet Project. Retrieved from: <http://www.pewinternet.org/2012/03/09/search-engine-use-2012/>
- Robots.txt (2007). About. Retrieved from: <http://www.robotstxt.org/robotstxt.html>
- Sherman, C. & Price, G. (2003). The invisible Web: Uncovering sources search engines can't see. *Library Trends*, 52(2), 282-298.
- Wright, A. (2009, February 22). Exploring a 'deep web' that Google can't grasp. *The New York Times*. Retrieved from: <http://www.nytimes.com/2009/02/23/technology/internet/23search.html?pagewanted=1&th&emc=th>