The Spies' Guide to Cyberspace
The NSA's Untangling the Web

Kevin O'Kelly, Guest Columnist

***Kevin O'Kelly** is Reference and Community Languages Librarian at the Somerville (MA) Public Library.*

*Correspondence concerning this column should be addressed to **Barry Trott**, Editor,* RUSQ, *Williamsburg Regional Library, 7770 Croaker Rd., Williamsburg, VA, 23188; email: btrott@wrl.org.*

After the news broke about the National Security Agency's (NSA) eavesdropping tactics, I read the agency's declassified document *Untangling the Web: A Guide to Internet Research*,[1] and remembered a scene from *The X-Files*. Mulder explains to Scully his theory of some mind-bogglingly nefarious government conspiracy. Scully looks at him patiently and says, "Mulder, you're talking about people who can't even balance the budget."

Listening to the latest news on the NSA, I feel like Mulder: these people are capable of anything. Reading the initial sections of *Untangling the Web*, I feel like Scully, and ask myself, *who are these people?*

I was shocked at how embarrassingly basic some of the tips are ("Browsers assume the prefix 'http://' unless you tell them otherwise" [p. 28]). Others are painful clichés ("When you do a search, you are going through more information in less than 30 seconds than a librarian could probably scan in an entire career thirty years ago" [p. 12]). The aforementioned are the sort of facts included in computer literacy classes for senior citizens. The argument can be made that a guide to Internet research needs to be comprehensive and cover all the basics. But how basic does one need to get? Apparently quite basic, although I can't imagine for whom could the brief explanatory section titled "Why Do We Need to Use the Internet?" conceivably be necessary.

This section of *Untangling the Web* illustrates the painful truth that most Internet users think they are better searchers than they actually are, and the authors are simply doing what good reference librarians should always do: adjust their instructions and assistance to the level of guidance the patron needs. Furthermore, that the authors felt the need to write about the web at such a basic level illustrates the ubiquity of digital illiteracy. I've had to explain to supposed "digital natives" the difference between an email address and a website URL. Apparently, even in the inner rooms of the Puzzle Palace, digital illiteracy abounds.

Much of the actual search advice is less embarrassing but nothing that a decent reference librarian doesn't already know, thus making an unintended argument for the continued necessity of librarians in the age of the search engine—even at the NSA. For example, use more than one search engine; depending on the type of information you want, a database may be a better source than Google; and—my favorite—"consider the accuracy and currency of information before using it" (p. 26). Brilliant. Anyone who reads *Untangling the Web* with the hope of learning how to spy on the neighbors—a hope encouraged by silly and inaccurate news coverage (e.g., from Wired.com, "Use These Secret NSA

Google Search Tips to Become Your Own Spy Agency")[2]—will be gravely disappointed, for the most part. The fact that a Freedom of Information Act request was necessary to make *Untangling the Web* public is a telling indication of the federal government's post–September 11 mania for secrecy and the ludicrous lengths it can reach.

After reading *Untangling the Web*, I can be sure of one thing: I missed a possible career path. Liberal arts majors frequently wonder what they can do for a living postcollege (yours truly was a history and English double major), and the authors of *Untangling the Web* probably have bachelor's degrees in English or comparative literature.[3] In the first twenty pages, I encountered references to Borges, Freud, Karl Popper, the myth of Sisyphus, Daedalus and Proteus, the tenth-century bibliophile Abdul Kassem Ismail, and *Tristram Shandy*. So, English majors of America: along with going to law school, getting an MLS, or signing up for AmeriCorps, you can consider becoming a spy—or at least working with them.

Untangling the Web was written in 2007 and is at times touchingly dated (at least to IT nerds). The author predicts that Ask.com has a shot at giving Google serious competition and recommends the now-defunct Pandia for metasearching. Caveats are necessary for other recommendations. I found the results called up by the metasearch engine Gigablast interesting but not exactly what I was looking for, regardless of the subject. For example, I did a search for "testing effect" just to see what I would find, hoping to share some of it with an education student I had been helping. The initial results were definitions (too basic for her needs), dissertation abstracts far too specialized for her needs, an eHow.com article on the "pre-testing effect" (psychological and emotional distress of people waiting to take a polygraph test), and a website on genetic testing. And while the search engine Exalead has impressive features (similar to Google), it seems to be most useful when the desired information can be found on Francophone websites.

All that having been said, I did learn quite a bit from the more advanced sections of *Untangling the Web*. The authors' praise of the unfortunately named metasearch engine Dogpile was spot-on—when I did a search for "testing effect" to see what it might come up with for my education student patron, the results included full-text articles and ERIC abstracts that matched her area of interest perfectly. I am not implying that Dogpile is inherently superior to Gigablast. My experience simply bears out the authors' advice that it is wise to use more than one search engine when looking for information. Reference librarians who use Google and only Google, take note. And embarrassingly, *Untangling the Web* helped bring me up to speed on Yahoo!, which hasn't been on my radar since the late 1990s. I was unaware that, until the 2009 deal in which Bing became the power behind Yahoo! Search, Yahoo! had temporarily become a unique search engine again. But more importantly, I rediscovered the value of Yahoo!'s human-selected resource directories. When researching topics of longtime interest, the links in certain Yahoo subcategories were a welcome change from repetitive Google hits. Because of the volume of websites tracked by its spiders, Google will be the search engine of choice for the foreseeable future, but it's worth remembering that the human-compiled directories can often provide more relevant information, faster. Size still matters, just not all the time.

The best sections of *Untangling the Web* are on searching the Invisible Web and on Google hacks (i.e., creative use of advanced search features). The Invisible Web,[4] for readers unfamiliar with the concept, is the information on the Internet inaccessible via conventional search engines, e.g., Bing and Google. In other words, almost all the Internet: database contents, dynamically generated pages intended for single use, webpages excluded by their creators—but basically just think databases, which includes everything from AcademicOneFile to the Internet Archive's WayBack Machine. According to the latest estimates, only 0.03 percent of the information on the Internet is available via search engine.[5] Although some of the information is inevitably dated, the authors were knowledgeable about the best Invisible Web resources available at the time of writing—The Wayback Machine (http://archive.org/web), Wolfram Alpha (www.wolframalpha.com), Infomine (http://infomine.ucr.edu), UC-Berkeley's Finding Information on the Internet: A Tutorial (www.lib.berkeley.edu/TeachingLib/Guides/Internet/FindInfo.html), and Phil Bradley's Making the Net Easier (www.philb.com), to name a few. And while all those sources are still around and are quite useful, I prefer Open Education Database's "Ultimate Guide to the Invisible Web" (http://oedb.org/ilibrarian/invisible-web) for a conceptual overview and Purdue's Online Writing Lab's "Resources to Search the Invisible Web" (https://owl.english.purdue.edu/owl/resource/558/07) as a starting point for searches.

I consider myself an old hand at Google hacking, and frankly, I was stunned by what is accessible online when I used *Untangling the Web*'s recommended search strategies. In this section, one of the authors coyly notes, "Nothing I am going to describe to you is illegal, nor does it in any way involve accessing unauthorized data. 'Google (or search engine) hacking' involves using publicly available search engines to access publicly available information that *almost certainly was not intended for public distribution*" (p. 175, emphasis mine). Among their tips: use the "filetype" command to limit a search to Excel spreadsheets and the keyword "login" to find lists of usernames and passwords, e.g. "filetype:xls login," or (for example) to limit your search further to Excel spreadsheets at Indian websites, "filetype:xls site:in login." Other suggestions included using keyword searches such as "proprietary," "confidential," or "not for distribution."

Among the information I retrieved using these search recommendations were the usernames and passwords of fifty-eight graduate students at Imperial College London, an Australian oil company's salary spreadsheet (labeled "confidential"), and a PDF of a Canadian organization's "strategic review initiative" marked "For Internal Distribution Only."

These sorts of searches are obviously unethical, and neither public nor academic librarians are likely to be asked for

help locating other people's usernames and passwords or corporate proprietary information. However, it does raise a basic question I ask myself frequently in my work as a public librarian: just because I can find information, should I? Patrons have asked me for help finding people's addresses and phone numbers. On one occasion, I was asked for help finding the not–readily available email address for an Israeli nuclear scientist. When I have conferred with colleagues on the ethics of these sorts of searches, reactions have ranged from, "It's our jobs to find information patrons want" to "You could be aiding and abetting stalking or giving someone up to a collection agency." There is no clearly "right thing to do" in most of these cases. *(Editor's note: O'Kelly's points here are an indication that the digital nature of information has only expanded the debate in the library world about the ethical implications of reference work, both in terms of access to information and to services provided. Readers interested in exploring this debate further may wish to see Jean Preer,* Library Ethics *[2008] and Robert Hauptman,* Ethics and Librarianship *[2002].)*

That said, the section on "Google hacking" (pp. 175–85) is worth reading for the simple reason that it lists numerous tools that can be used to focus a search, and frankly most of us tend to go the lazy route, opting for Googling keywords and not taking advantage of the ways we can separate the virtual wheat from the chaff.

The final section of *Untangling the Web* is titled "Internet Privacy and Security—Making Yourself Less Vulnerable in a Dangerous World." Without the slightest trace of irony, the authors cover the basics of Internet privacy, including browser settings and preferences (for both Internet Explorer and Firefox—this entire document is PC-centric), the advantages of emailing offline, the perils of autocomplete, and the security vulnerabilities of JavaScript. And, in a return to the surreal basics of *Untangling the Web*'s opening section, they seem to think it necessary to explain what phishing is and to warn employees of one of the world's largest domestic intelligence agencies that they shouldn't download a program or open an attachment unless they know what it is. However—snark aside—this material is all worth sharing with patrons, given how innocent most Internet users are about security issues. I'm simply surprised it was all included in an in-house manual for spooks.

It's worth every public and academic librarian's while to read at least parts *of Untangling the Web*. Despite being six years old, it's still an excellent primer on beefing up online research skills and a much-needed reminder of the myriad vulnerabilities that are an inevitable part of using the Internet.

References

1. The Center for Digital Content, *Untangling the Web: A Guide to Internet Research*, updated February 28, 2007, accessed December 3, 2013, www.nsa.gov/public_info/_files/Untangling_the_Web.pdf.
2. Kim Zetter, "Use These Secret NSA Google Search Tips to Become Your Own Spy Agency," Wired.com, May 8, 2013, www.wired.com/threatlevel/2013/05/nsa-manual-on-hacking-internet.
3. Their names have been redacted from the declassified document but are listed on the original FOIA.
4. AKA the Deep Web, the Dark Web, the Undernet, and the Hidden Web.
5. "The Ultimate Guide to the Invisible Web," Open Education Database, November 11, 2013, accessed December 3, 2013, http://oedb.org/ilibrarian/invisible-web.