The Librarian as Hacker, Getting More from Google

R. Philip Reynolds
Research Education Librarian
Stephen F. Austin State University

Abstract

This paper will cover four areas. First it will discuss the research habits of search engine users and some of the problems with these habits. Then it will discuss librarians' use of search engines. Here we encounter the real question: Do we do much better? Can we use a search engines to their full potential? When needed, can we hack an engine to make it perform beyond its intended function? Can we use a clever workaround to solve a problem? Or are we on a level playing field with our patrons once we get outside traditional database searching? Google currently offers over seventy free services—fifty-two of them are search related. How many are we familiar with and comfortable using?

The paper will incorporate a discussion of some of the Google hacks documented in the book Google Hacks, 3rd edition, by Paul Bausch, Tara Calishain, and Rael Dornfest. It will conclude with a demonstration of how to use Google to create Custom Search Engines (CSE) that can be used to support curriculum or enhance a collection of primary or other specific sources.

Introduction

Traditionally hacking has been associated with illegal activity. There is a positive definition to hacking. A "hack" can mean a clever solution that solves problems. Hacks are unorthodox solutions that extend the capability of an application beyond its conventional or intended use. Librarians can use or invent hacks to get more from their databases and applications. This paper will demonstrate ways to extend or hack Google to go beyond its conventional uses.

What does hacking have to do with librarianship?

One area where librarians need to come up with clever solutions or unorthodox ways of doing things is in searching for information on the Internet. Anybody can type in a couple of keywords and do a search, but to really stretch Google's capabilities we need to be able to do a hack. This requires in-depth knowledge of Google and some understanding of how web pages are organized and built.

Resource discovery is an obvious use for a Google hack. Resource discovery may appear straightforward, but in reality requires skills. Unfortunately these skills do not prevail among our users. A recent study by the Pew Charitable Trust reveals how poor the research being done by search engine users really is:

- 92% of those who use search engines say they are confident about their searching abilities, with … 52%, saying they're "very confident".

- 87% of searchers say they have successful search experiences most of the time, including some 17% of users who say they always find the information for which they are looking.
- 68% of users say that search engines are a fair and unbiased source of information … (Fallows i)

Furthermore, Fallows reports that "97% of internet users under 30 years express confidence in their search skills" (5).

Think about those who have recently visited the reference desk. How competent were their search skills? This begs the question raised by Chris Sherman, executive editor of Search Engine Watch:

> What makes searchers so confident in their own abilities? "The majority are doing simple searches," said Deborah Fallows, Senior Research Fellow at the Pew Internet & American Life Project and author of the report. "It's very easy and very quick to get an answer for a passing thought, and that leads to confidence." (Sherman).

Librarians understand the problems with relying on the web for information. Large amounts of research have gone into looking at the structure of web pages and using that structure to more effectively mine information from the pages (Cohen 1). Even more effort has gone into encouraging webmasters to add metadata to their page's Meta tags. None of these efforts have turned into that which every librarian really desires and longs for; a catalog for the entire web.

This new environment that we did not create, do not control, and can not classify or catalog is where our customers went. In a recent study, "…it became apparent that students are very eager to use only the Internet in conducting research. Though the survey was not in any way limited to Internet resources, less than 2% of students' responses to all questions included non-Internet sources." Up to 75% of the students in the study relied on the first answer they found and did no further research (Graham 72, 74). Most searches are just two or three key words. The average searcher checks less than two pages of search results (Fallows 2). Even when seeking medical information, 75% of those surveyed said they rarely or never check the source or the date of the health information they find on the web (Fox iii).

Do we do much better? As librarians are our search engine skills as highly developed as our OPAC or database skills? When needed, can we hack the engine to make it "perform beyond its intended function" or can we apply that "clever solution or workaround that solves" the problem? "A recent comparison of Cornell University reference librarians and Internet users on Google Answers showed reference librarians with their vastly larger collection of quality print and electronic information, years of experience, and professional training scored little better than the Internet users offering information on Google Answers. The researchers seem to try to excuse these results (Reynolds) by saying "A final point on the evaluations involves sources. Google researchers (as opposed to the librarians) are experts at locating hard-to-find information on the Web. Their answers, therefore, tend to be limited to freely available networked resources" (Kenny 10, 14).

What Can We Do?

"It's time to learn the ins and outs of search engines" and do some hacking (Reynolds). Many search engines accept Boolean operators. Google also allows certain syntax in place of them. The pipe symbol | can be used for OR; the – symbol can be used for not. Google assumes the operator AND for each term in a search. For example [dog AND bark] is the same as [dog bark]. The + sign is intended to force Google to include a term or character it would normally ignore. For example, the terms Sam I am would probably return a lot of web pages with the word Sam, but Google would ignore I am. We can force them to be included in the search like this: [Sam +I +am]. What we really want is the phrase Sam I am. This is searched with quotation marks ["Sam I am"] (Bausch, Calishain, and Dornfest ch.1, sec.1).

Google implements automatic stemming technology as a part of their searches and uses the asterisk as a full word wildcard ("Google Help Center: The Essentials of Google Search"). It can be used to search for a phrase that one does not quite remember. For instance, ["The mass of men live lives of * desperation (Bausch, Calishain, and Dornfest ch.1, sec. 2). The wildcard can return results with one word or with many words in its place. According to Google: "…a search for [cooking * classes] will match the phrases 'cooking school classes' and 'cooking and wine tasting classes'" ("Does Google Support Wildcard Searches"). Another unique Google syntax is the tilde ~, allowing one to search for pseudo-synonyms. Google finds words one would expect to find in a traditional thesaurus, but it also turns up unexpected terms, because Google uses algorithms to identify synonyms instead of human philologists.

Web page metadata can be exploited by savvy searchers. The <Title> tag of a web document often describes either the subject of the page's content, or the department or institution where it originated. The URL describes the page's domain. Less familiar than the ubiquitous [.edu] or [.com] are: [.us] from state or local government, domains from other countries [.de] (Germany). The directories listed in the URL tell us things like how important the page is, what function it serves or possibly what it contains. This may seem obvious to most that have had exposure to HTML, but when we use this metadata in an unorthodox search, we can discover valuable resources. This metadata is accessed by the use of special syntax available in Google. In Google, domain information is accessed with the [site:] syntax, title information with [title:] or [allintitle:] syntax, and URL information with [inurl:] or [allinurl:] syntax (Bausch, Calishain, and Dornfest ch.1, sec. 3).

We can take advantage of more than one type of metadata in the same search, resulting in powerful and sophisticated hacks (Bausch, Calishain, and Dornfest ch.1, sec. 4). Searching Google and with the following search string produces amazing results.

site:mil (intitle:database | inurl:database)

Let's examine our search. The first operator is [site:mil]. Site delineates the domain. Next we have a colon immediately followed by the domain name [mil]. This tells Google to search only in military web sites. It will use metadata from the domain name on each web page to sort out the military publications. Next there is a space. When Google finds a space between terms it

assumes the operator [AND]. Then we have parenthesis which sets up the next half of our search. The operator [intitle:] tells Google that the next term must appear in the <Title></Title> tag in the HTML code. Again this can be a form of metadata identifying what a page contains or its subject. Next there is a colon and immediately following it the word [database]. Whatever page Google retrieves, it must have the word “database” in the <Title></Title> tag.

The search string allows for an alternative or additional search of metadata by adding a space and then the pipe symbol [|]. The pipe symbol is borrowed from many programming languages and is the same as the [OR] operator. Next comes the operator [inurl:] followed by the word [database] again. This tells Google that if the word “database” appears anywhere in the URL, even at the end of several directories and slashes, that it is to retrieve it. Our search is going to find all pages with the “<title> database </title>” in the HTML code or the word “/database/” in the URL on military web sites. The results of the search are shown below: (see fig. 1)

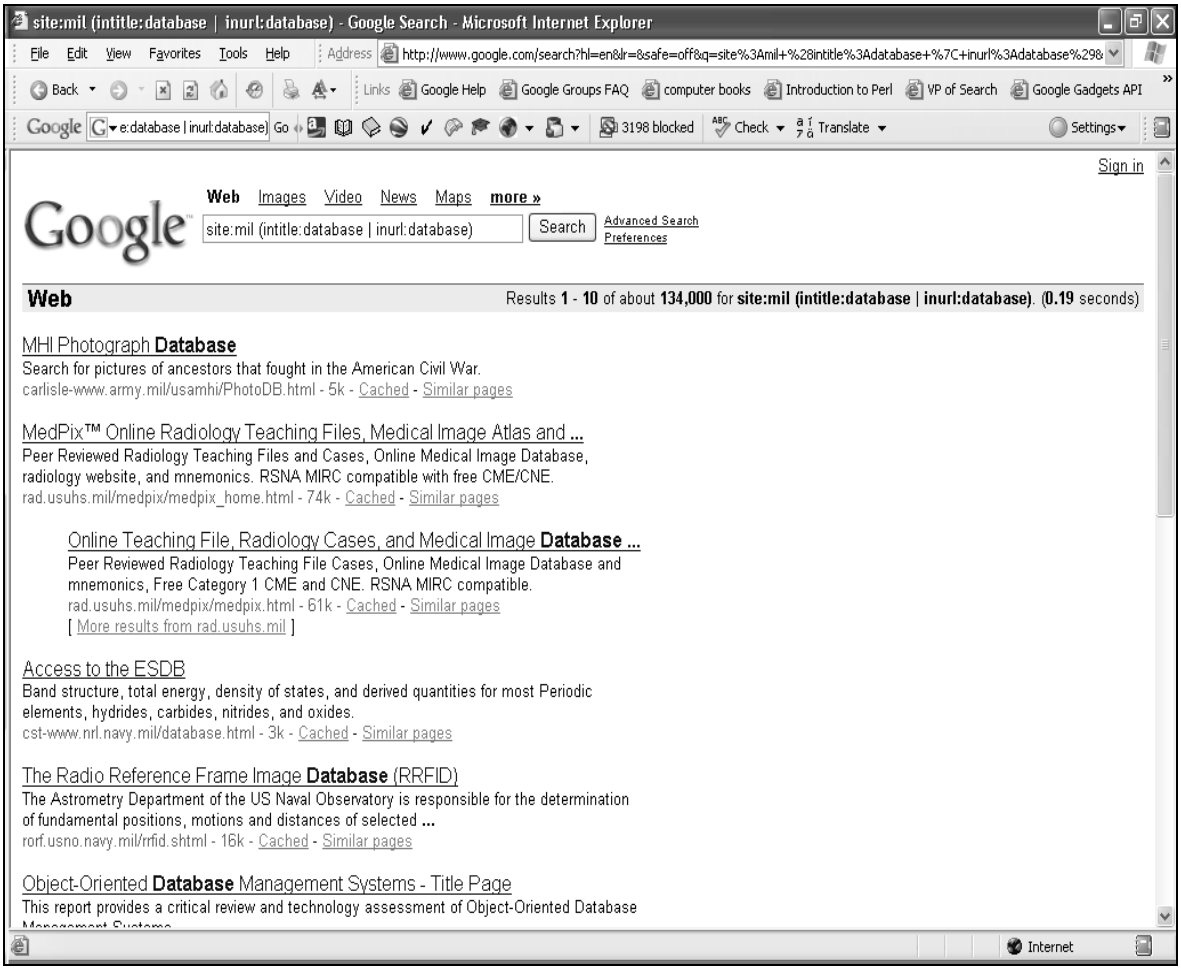

Fig. 1. The results of the Google search site:mil (intitle:database | inurl:database).

Our first hit is a Civil War photograph database. Our next two hits are to a database called MedPix. Looking at one of these links we find:

> **MedPix™** is a **free** online **Medical Image Database**, provided by the Departments of Radiology and Biomedical Informatics, Uniformed Services University, Bethesda, MD. All public content is **peer-reviewed** by an Editorial Panel (MedPix).

Next there is the ESDB database. The link to this database is dead. Going to the home page reveals the Center for Computational Materials Science, a Naval Research Laboratory. The page provides access to a wealth of research information, including many valuable databases. The list of primary and research resources uncovered by this search goes on and on.

We have created a virtual collection. The search string is the collection development statement. The materials collected are the web pages and databases found in the search results. The collection is virtual because it does not exist as a collection anywhere. It comes into existence at the time of the search and is gone when the search page is closed. Collection bias results from the choice of search engine and the algorithms used by that search engine. With a few minor modifications other virtual collections can be created (see table 1).

Table 1
Google Searches that Create Virtual Collections

Search	**Type of Results**
site:gov (intitle:database | inurl:database)	valuable federal government databases
site:edu (intitle:database | inurl:database)	databases at universities
site:il.us (intitle:database | inurl:database)	databases on Illinois state government web sites

We can repeat this search with any domain (.org, .net, .int, .uk, etc.) and refine our results to specific types of information by adding keywords for the subject we are interested in such as crime, plants, medicine, etc.

What Can Google Really Do?

We need to be able to find the information that patrons can not find for themselves. This entails being able to stretch the capabilities of a search engine beyond its normal limits (Bausch, Calishain, and Dornfest,; Johnny). If we count all of the Google services listed under "More Google Products" and "Google Labs," there are seventy different products or services. Fifty-two of those are search related (Reynolds). In their book Google Hacks, Bausch, Calishain [and] Dornfest list hack #1 as being aware of and using the Google directory, #2 as Google Zeitgeist, #4 as the spellchecker, and #5 as the Google Phonebook (ch. 2, sec. 1, 2, 4, 5). By studying the documentation and tools provided by Google, we immediately become expert searchers (Reynolds). An easy way to incorporate this knowledge is to remember the three P's of becoming a search engine expert: "**Pick** two or three search engines to use regularly. **Print** and study any help pages or documentation they provide. **Practice** using the various tools and incorporate them into your own research and into your work with others" (Reynolds).

If we look at the search operators available through Google, we find many of pieces of metadata we can use in searching. Many can be found on the Google Help: Cheat Sheet, Google Help

Search Center: Advanced Operators: Alternate Query Types, Google Help Search Center: The Essentials of Google Search and on the help screen for the various special searches like Google Scholar Help: Understanding a Search Result. If we add up the different syntax or operators available for a regular Google search listed on the help pages, we come up with over 28 different operators. Each of these operators focus on a specific type of metadata found on the web. There are other operators not mentioned on these pages, such as [filetype:], [daterange:] and others that appear to have little or no documentation at all (Bausch, Calishain, and Dornfest ch. 1, sec. 3).

Something New

Google has released a "Custom Search Engine (CSE)." It can be found on "Google Co-op: Welcome to Google Co-op." To use it, we must create a Google Account. You will be prompted to create one when you select the "Create a Search Engine" button on "Custom Search Engine." Once you have an account, Google allows you to create your CSE. However, what is this engine going to search? This crucial question that will make the difference between just another web page and a quality research tool.

I recently noticed an increase in book sales in WWI history. I thought it would be great if we had a collection of first hand accounts from soldiers about the war. If the same type of collections could be created for other wars, comparisons could be made between soldiers' opinions, attitudes and experiences. I could create a virtual collection of soldiers' diaries and journals from the war that had been placed on the web. Most "virtual collections" are a series of image files inside of proprietary databases. This makes it impossible to do a full text search for information, such as what was the food like or what was the first poison gas attack like. Even if there are transcripts of the image files in the database, they are still inaccessible to search engines because of the proprietary nature of most database systems.

This virtual library would need to collect diaries and memoirs of WWI soldiers that search spiders could crawl. These needed to be in established web sites that were unlikely to disappear and that seemed to post accurate transcripts of the original documents. Armed with these criteria I began searching for the diaries. I started with the web directories and then used some of the search strategies discussed earlier.

(WWI OR "The Great War" OR "World War One") (intitle:diary OR intitle:diaries OR inurl:diary OR inurl:diaries)

With this and other searches, I collected a list of resources with documents that matched my collection criteria.

The "Create a Custom Search Engine" page asks for the new search engine's name, then a description. Keywords are added to help others find the search engine. Terms like "WW I, World War One, The Great War, War, Diaries, Soldiers" worked well. Google provides an entry box in the creation form to list the sites to search. The CSEs tool has its own set of syntax. A chart describes the patterns and syntax used to enter the URLs in the documentation section. I spent over an hour entering my data. When I went to another web page to check something and then went back, all of my data was lost. The next time I entered my data, I created a plain text file in

"Windows Notebook" and saved it after every entry. After the text file was complete, I copied and pasted it into Google's form. Some of the sites had all of their diaries in one directory or part of the site which made it very easy to enter the data. Brigham Young University Library had all of their journals in: http://www.lib.byu.edu/~rdh/wwi/memoir/*.

Earlier the asterisk acted as a wildcard for a word. However, in the CSE it is used to include sections of sites without naming every single page ("URL Patterns"). On some sites the asterisk worked well; others had to have the entire URL for each page listed to avoid unrelated documents in the same directory or in related directories. After all the data was entered and saved, Google created the search engine and a free homepage to host it on. I have created or started several CSEs. The WW I: Diaries of The Great War engine is at:

http://www.google.com/coop/cse?cx=014997348189176913774%3A_b4nqqxhk2k

A second search engine Sacred Text of Islam The Quran and Hadith is at:

http://www.google.com/coop/cse?cx=014997348189176913774%3Abovrnnvevwo

The others can be found at:

http://www.google.com/coop/profile?user=01499734818917691377 4

One of the features of the CSE is that you can take the code of your engine and put it on your own web site along with the results page. When you are logged into your account, your homepage has a link labeled "Edit this search engine" When you click on that link, you are brought to a page where you can edit all facets of your engine. One of the options is labeled "code." If you select that option, Google will provide you with the necessary source code to add the search engine to your library's site or your subject page or any other site you have. There is an example of the WWI engine at:

http://www.protoknowledge.com/iwasthere/war/worldwarone/index.html

The search box can be integrated into your site and, if you try a search, you will see that the results page also has the look and feel of the rest of the site.

Conclusion

Google and other search engines have become a part of the information landscape. In order for us to continue to claim the status of information professionals, we will have to add "Google Hacks" to our search skills. We need to be able to not only fully exploit the Internet, but we must also use our knowledge and experience to create tools like CSEs and any others that may prove useful. We do this with the hope that we will help our patrons and expand their research skills, options, and capabilities.

Works Cited

Bausch, Paul, Tara Calishain, and Rael Dornfest. Google Hacks. 3rd ed. Sebastopol, CA: O'Reilly Media, Inc., 2006. Safari Books Online, Ralph W. Steen Library, Nacogdoches, TX.

Cohen, William W. "Recognizing structure in Web pages using similarity queries." Proceedings of the Sixteenth National Conference on Artificial intelligence and the Eleventh innovative Applications of Artificial intelligence Conference innovative Applications of Artificial intelligence. Orlando, FL, Menlo Park, CA: American Association for Artificial Intelligence, 1999.

Fallows, Deborah. "Search Engine Users: Internet Searchers are Confident, Satisfied and Trusting – But They are Also Unaware and Naïve." Pew Internet & American Life Project: Reports Online Activities & Pursuits. 23 Jan. 2005. Pugh Research Center. 04 Nov. 2006 <http://www.pewinternet.org/pdfs/PIP_Searchengine_users.pdf>.

Fallows, Deborah, Lee Rainie, and Graham Mudd. "The Use of Search Engines is a Top Online Activity and Americans Increasingly Feel They Get the Information They Want When They Perform Search Queries." Pew Internet & American Life Project. Aug. 2004 1-7. 06 Nov. 2006 <http://www.pewinternet.org/pdfs/PIP_Data_Memo_Searchengines.pdf>.

Fox, Susannah. "Online Health Search 2006: Most Internet Users Start at a Search Engine." Pew Internet & American Life Project. 29 Oct 2006. 06 Nov. 2006 <http://www.pewinternet.org/pdfs/PIP_Online_Health_2006.pdf>.

"Google Custom Search: Custom Search Engines." Google. 2006. Google, Inc. 6 Nov. 2006 <http://www.google.com/coop/manage/cse/create/1?>.

"Google Co-op: Custom Search Engine: URL Patterns." Google. 2006. Google, Inc. 6 Nov. 2006 <http://www.google.com/coop/docs/cse/patterns.html>.

"Google Custom Search: Custom Search Engine." Google. 2006. Google, Inc. 6 Nov. 2006 <http://www.google.com/coop/cse/overview>.

"Google Frequently Asked Questions - File Types." Google. 2005. Google, Inc. 6 Nov. 2006 <http://www.google.com/help/faq_filetypes.html>.

"Google Help Search Center: Advanced Operators: Alternate Query Types." Google. 2005. Google, Inc. 6 Nov. 2006 <http://www.google.com/help/operators.html>.

"Google Web Help Search Center: Does Google Support Wildcard Searches?" Google. 2007. Google, Inc. 15 Aug. 2007 < http://www.google.com/support/bin/answer.py?answer=3178&topic=352>.

"Google Help Search Center: The Essentials of Google Search." Google. 2005. Google, Inc. 6 Nov. 2006 <http://www.google.com/help/basics.html>.

"Google Help: Cheat Sheet." Google. 2005. Google, Inc. 6 Nov. 2006 <http://www.google.com/help/cheatsheet.html>.

"Google Scholar Help: Understanding a Search Result." Google. 2006. Google, Inc. 6 Nov 2006 <http://scholar.google.com/scholar/help.html>.

"More Google Products." Google. 2006. Google, Inc. 6 Nov. 2006 <http://www.google.com/intl/en/options/>.

"Google Co-op: Welcome to Google Co-op." Google. 2006. Google, Inc. 6 Nov. 2006 <http://www.google.com/coop/>.

"Google Labs." Google. 2007. Google, Inc. 15 Aug. 2007. <http://labs.google.com/>.

Graham, Leah and Panagiotis Takis Metaxas. "'Of course It's True; I Saw it on the Internet!': Critical Thinking in the Internet Era." Communications of the ACM 46.5 (2003): 70-75.

Generic Top-Level Domains. 21 Dec. 2005. IANA: Internet Assigned Numbers Authority. 28 Nov. 2006 <http://www.iana.org/gtld/gtld.htm>

Root-Zone Whois Information: Index by TLD Code. 26 Nov. 2006. IANA: Internet Assigned Numbers Authority. 28 Nov. 2006 <http://www.iana.org/root-whois/index.html>

Kenney, Anne R., Nancy Y. McGovern, Ida T. Martinez, and Lance J. Heidig. "Google Meets eBay: What Academic Librarians Can Learn from Alternative Information Providers." D-Lib Magazine 9.6 (2003) 1-16. 06 Nov. 2006 <http://www.dlib.org/dlib/june03/kenney/06kenney.html>.

Long, Johnny. johnny.ihackstuff.com. 06 Nov. 2006. 10 Nov. 2006 <http://johnny.ihackstuff.com/index.php?module=prodreviews>.

MedPix, "MedPix Medical Image Database." MedPix. 2006. Departments of Radiology and Biomedical Informatics, Uniformed Services University. 17 Oct. 2006 <http://rad.usuhs.mil/medpix/medpix_home.html>.

Naval Research Laboratory, "Center for Computational Materials Science Code 6390." Materials Science and Technology Division, Naval Research Laboratory. 11 June 2002. United States Navy. 6 Nov. 2006 <http://cst-www.nrl.navy.mil/>.

Reynolds, R. Philip. "Things to Read Before Session." Electronic Resources & Libraries. 7 Feb. 2007. ER& L Forum. 14 Aug. 2007 <http://electroniclibrarian.org/forum/2007/02/07/things-to-read-before-session/>.

Sherman, Chris. "Survey: Searchers are Confident, Satisfied & Clueless." SearchDay. 24 Jan. 2005. Search Engine Watch. 06 Nov. 2006 <http://searchenginewatch.com/showPage.html?page=3462911>.