Harnessing the Power of Google

What Every Researcher Should Know

Christopher C. Brown

HARNESSING THE POWER OF GOOGLE

What Every Researcher Should Know

Christopher C. Brown

LIBRARIES UNLIMITED™
An Imprint of ABC-CLIO, LLC
Santa Barbara, California • Denver, Colorado

Library of Congress Cataloging-in-Publication Data

Names: Brown, Christopher C., 1953– author.
Title: Harnessing the power of Google : what every researcher should know / Christopher C. Brown.
Description: Santa Barbara, California : Libraries Unlimited, an imprint of ABC-CLIO, LLC, [2017] | Includes index.
Identifiers: LCCN 2017007461 (print) | LCCN 2017021554 (ebook) | ISBN 9781440857133 (ebook) | ISBN 9781440857126 (acid-free paper)
Subjects: LCSH: Google. | Internet searching. | Internet research. | Libraries and the Internet. | Internet in library reference services.
Classification: LCC ZA4234.G64 (ebook) | LCC ZA4234.G64 B76 2017 (print) | DDC 025.04252—dc23
LC record available at https://lccn.loc.gov/2017007461

ISBN: 978-1-4408-5712-6
EISBN: 978-1-4408-5713-3

21 20 19 18 17 1 2 3 4 5

This book is also available as an eBook.

Libraries Unlimited
An Imprint of ABC-CLIO, LLC

ABC-CLIO, LLC
130 Cremona Drive, P.O. Box 1911
Santa Barbara, California 93116-1911
www.abc-clio.com

This book is printed on acid-free paper ∞

Manufactured in the United States of America

Contents

This page intentionally left blank

Illustrations

FIGURES

TABLES

Introduction

Why is a reference librarian writing about using Google for academic research? Don't professors tell students *not* to use Google in their research? Isn't Google a threat to librarianship, and won't it eventually replace the need for librarians? This book will suggest that Google is extremely valuable in the academic research process, but users need to understand what is being searched, how to constrain searches to academically relevant resources, how to evaluate what is found on the Web, and how to cite what is found.

It's not uncommon for new university students to think they already know how to search. Their first paper comes due, and what do they do? They resort to using Google. They think they know it all—or at least where everything can be found. But when they get that first paper back with an unsatisfactory grade and comments like, "you need to cite peer-reviewed articles," "don't rely on Google," and "you need reliable sources to support your arguments," many of them show up for research consultations and to meet with reference (research) librarians. It takes a librarian to really show them how to search.

I'll let you in on a little secret: all reference librarians, academic or otherwise, use search engines, especially Google. The extent to which it is used varies, of course, but Google can be the single best starting point for navigation down the right path. When someone has a question, they don't know the answer. This seems obvious, but it is profoundly important. If someone doesn't know something, they may not even know how to visualize what the answer looks like or what path to pursue. Trying to navigate in a fog is nearly impossible. But Google is there to correct our misspellings, suggest new pathways, and clear the fog.

This book is not intended to cover every feature of Google. We intentionally gloss over Google Earth, most of the Google widgets, Google personalization features, linkages to Google+ and other Google properties, and even some of the search capabilities that have little to do with academic research. This is intentional. Many books already do that. All you have to do is search Google Web like this: *how to search google*, or Google Books like this: *how to search google*. This book is focused on assisting students, researchers, teachers, professors, and librarians in finding primary and secondary sources using Google.

1

Searching Generally

The history of language, writing, and indexing is a fascinating one. When thinking of how we access the vast libraries of information that exist in the world, we must remember that before the writing of literate cultures there were oral cultures. These cultures had ways of remembering or indexing in the mind as well. Long gone are the days when scholars would memorize long texts and be able to search their memories for ideas and arguments. Since Gutenberg's printing press and the printing revolution, publications have proliferated and various finding aids, including bibliographies, printed indexes and catalogs, card catalogs, and, more recently, online catalogs and indexes, have been created to provide intellectual access to print publications. To fully appreciate the technology available to us today, we need a bit of historical perspective.

HISTORY OF SEARCHING

The blossoming of magazine and journal publishing soon necessitated a way to discover all this content. Thus modern indexing was born. A glance at a technology timeline will help give us a historical perspective (see Figure 1.1). The mid-1800s saw the beginning of periodical indexing with pioneers like William Frederick Poole and H. W. Wilson. Poole's *Index to Periodical Literature*, published in the mid-1800s through various editions and supplements, economized space with abbreviations and small print and was tedious and challenging to use, but it worked. Wilson, whose work endures to this day, also saved space with abbreviations, but incorporated a technology that was being developed in his day: subject headings along the lines of Library of Congress subject headings.

Some of our older readers will remember libraries with card catalogs—those handsome wooden cabinets with tiny drawers to accommodate cards

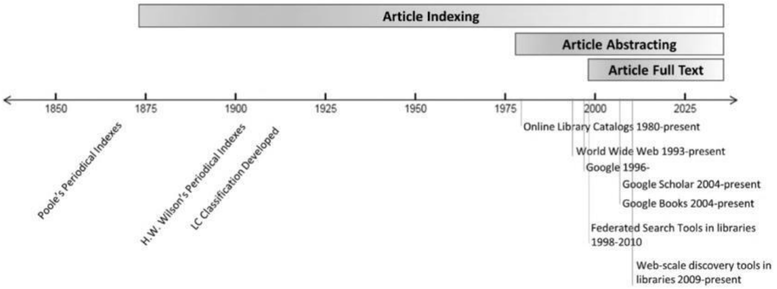

Article Indexing
Article Abstracting
Article Full Text
1850
1875
1900
1925
1950
1975
2000
2025
Poole's Periodical Indexes
H.W. Wilson's Periodical Indexes
LC Classification Developed
Online Library Catalogs 1980-present
World Wide Web 1993-present
Google 1996-
Google Scholar 2004-present
Google Books 2004-present
Federated Search Tools in libraries 1998-2010
Web-scale discovery tools in libraries 2009-present

FIGURE 1.1. Key Events in Recent Searching History.

with information about the books or other materials owned by the library. The most common scheme for library card searching was the dictionary catalog approach: cards arranged alphabetically for authors, cards for subjects, and cards for titles. But it gets more complex that just these three simple categories. When there are multiple authors or editors, another card set needs to be created. Every subject assigned generates more card sets. Title cards would account not only for the main title, but also for additional titles such as series titles, varying forms of the title, translated title, etc.

There was no keyword searching in the physical library catalog days. Access, at least for English language materials, was "left-anchored." That is, searchers had to start from the left to look up an author, subject, or title. If an author's name was Christopher C. Brown, the name was inverted "phonebook" style to Brown, Christopher C. Subjects were governed with controlled vocabularies such as the Library of Congress subject headings. Titles had special considerations as well. Users had to omit initial articles (for English, omitting *a*, *an*, or *the* from the beginning of titles). This greatly inhibited the access to materials, but because that was the state of the art at that time, users didn't know what the future held.

The late 1970s and 1980s saw the development of online catalogs. Because libraries were worried that the public would not accept the new technology, online catalog records were made to look like printed catalog cards. But there was one major advancement of technology that was transformative: the ability to search the online catalog record by keyword. In other words, users wouldn't need to think about inverting an author's name. If all that was known of a title was several words within the title and perhaps an author's first name, the book could likely be found. Users unaware of the proper formation of subject headings could nevertheless locate materials simply by searching using keywords. This technology was a huge forward leap and should be fully appreciated.

The quest for magazines or scholarly journal content experienced a transformation similar to books in library contexts. When indexes no longer had to be printed out, economy of space was less important. Abstracts of articles

TABLE 1.1. Contrast between E-books and E-journals.

E-books	E-journals
Entire books online, but very often with DRM restrictions.	Single articles available through publishers and aggregators, but no DRM.
Restricted printing and downloading.	No restrictions on printing and downloading.
Sometimes limits on "simultaneous users," making use for course-related materials challenging.	No simultaneous user restrictions.
Often requires establishing a login with the vendor.	Vendors generally do not know the identity of users.
Often requires downloading helper software with associated barriers.	Usually only requires Adobe Acrobat Reader, which is almost universally installed.

could be incorporated into the entry for each article. Keyword searching likewise transformed access to scholarly journal literature.

But the technology didn't stop there. Full texts of e-books and articles began to appear in the late 1990s and early 2000s. Early e-books were not fun to use. Digital rights management (DRM) systems made access onerous. In order to sell their new e-book model to print publishers, e-book vendors tried to replicate the print user experience with their digital books, making the argument that the online model completely replicated a print user model: one user per time, per e-book. But users didn't understand it that way: they wanted unlimited access to online content, not "one simultaneous user." Other vendors applied DRM with helper applications such as Adobe Digital Editions. Downloading of auxiliary software places additional barriers before the user, as evidenced by the many assistance calls placed to reference desks. These and more e-book barriers persist to this day.

Where e-books failed, e-journals succeeded (see Table 1.1). Books often had DRM protections, but journals didn't need this, because it was the part (individual articles) and not the whole (entire books) that were being exposed. The only barriers users had to e-journal content was whether their library subscribed or not and was the initial authentication process. Once users were authenticated, they could easily save entire e-journal articles, print them out, and read them either online or as a printout.

Google entered the world of full text of both e-books and e-journals with Google Books and Google Scholar, respectively. But more about those models later. Suffice it to say that these two additional Google initiatives transformed the way students think about content. From the initial search to finally accessing the full text, the discovery and fulfillment process was forever changed.

TENSION BETWEEN CONTROLLED VOCABULARY AND FULL TEXTUALITY

The "holy grail" of searching is to find "all and only" the relevant materials. How one goes about finding this "all and only" has been gradually shifting. Traditionally librarians have emphasized established search techniques when doing library instruction. These include Boolean search terms (AND, OR, NOT), nesting with parentheses, and proximity operators (these vary by database, but may look like w/10, NEAR5, or something similar). In addition instruction librarians spend much time teaching the mastery of searching by controlled vocabularies. Controlled vocabularies are agreed-upon vocabulary sets, established by subject experts using thesaurus construction standards (NISO 2010), for use within a specific discipline. For example Psychological Abstracts (and its online analog PsycInfo) are controlled by the *Thesaurus of Psychological Index Terms*; Sociological Abstracts by the *Thesaurus of Sociological Indexing Terms*; and ERIC, the U.S. Department of Education's Education Resources Information Center index and database, has as its thesaurus the *Thesaurus of ERIC Descriptors*. In science and engineering there was the INSPEC Thesaurus and the IEEE Thesaurus, among many others. The idea behind subject-specific thesauri was to capture the discipline-specific nuances of terminology and to apply it consistently within that discipline. It should be noted, however, that even between disciplines as close as education, psychology, and sociology there were sometimes great differences in assigned terminology and thus in the resulting application of those terms to indexed items.

For example, for the notion of e-mail, the ERIC descriptor is "Electronic Mail," psychology uses "Computer Mediated Communication," and sociology uses "Telecommunications" often combined with "Interpersonal Communication." Yet each of these official descriptors lags behind the culturally accepted term "e-mail." Indexing, in all of its structures and standards, is far from a perfect art and is often variously applied, depending on the indexer.

However, Google's searching power has become so popular with users—and with librarians—that advanced searching techniques and reliance on thesauri has decreased in recent years. Do we still need thesauri? Yes we do, especially in the fields of medicine and law. Do we still need Boolean operators and other connectors? Yes, because certain databases require them to produce predictable search results. Indexing and abstracting products (or as librarians refer to them, A&I products) are easily searched with Boolean operators (AND, OR, NOT). These operators work well because what was being searched was only the words contained in the index record: authors, title, subject, and perhaps an abstract—a relatively small set of words. But when searching full-text books, the AND and OR operators all of a sudden lose much of their power. Suppose you are searching for hummingbirds in

Colorado. Searching *hummingbirds AND Colorado* in an abstract record (metadata only) will easily uncover relevant materials, because the span of words searched is constrained to capture the true "aboutness" of the word being indexed. But when using the same Boolean operators in a database capable of searching across the full texts of thousands of millions of books, too many irrelevant materials are going to be retrieved. "Colorado" may appear on page 3 of a book, and "hummingbird" may appear on page 250. But is this book really about hummingbirds in the context of Colorado given that the two terms are so far removed from each other? To effectively perform a search across such large collections of full texts we need one of two things: 1) proximity operators that can constrain the closeness of one term to another, thus increasing the chances of a greater degree of relevancy, or 2) a sophisticated relevance ranking technology that is able to place at the top of the results list materials where two or more search terms are in close proximity to one another, where weighting is given to structural considerations like title, subdivisions, assigned descriptions, etc., and a sophisticated underlying system of synonymy is employed. The former system of proximity operators is what library vendors typically provide in their products as described deep within their help modules; the latter system is what search engines like Google typically do for us. Do users understand the differences between how to frame a search in Google versus how to put together a search in a typical library database? Do users know how to work proximity operators when searching a library-subscribed full-text database? The answers to these questions are likely "no." This helps us understand the popularity of Google Web, Google Scholar, and Google Books and helps explain why researchers often don't start their research with library resources.

The problem with Google is that Google isn't telling us exactly how the results are retrieved and how the relevancy is ordered. We simply see many more resources for the same search, and the ordering of results is mysteriously up to Google. I recommend that users do a "both-and" approach. Search both Google and traditional library databases. When I work with students in reference consultations, I very often take them to Google Scholar first, because they can accumulate a lot of relevant resources very quickly. Then I tell them to play "clean up" in the subject-specific library databases. This kind of joint strategy is especially important for doctoral students who are responsible for reading everything (not just some things) about their topics.

There are two disciplines where exactness in searching is absolutely essential: medicine and law. We don't want our doctors to miss crucial content about their field. Nor do we want attorneys with expensive "billable hours" to overlook materials that could help our legal standing as they represent our interests. In these disciplines there is heavy reliance on the application of controlled vocabularies. The National Library of Medicine has developed their *Medical Subject Headings* (MeSH) to capture disease names, medical procedures, and other medical terms. Staff at the National Library of Medicine with subject expertise carefully assign MeSH headings for precision in retrieval of results.

In the legal realm, the proprietary West Key Number System is a long-standing authority for indexing precedent-setting cases, law reviews, legal encyclopedias, and other materials. These legal subjects are assigned by attorney-editors who are specialists in various areas of law. It is important for legal researchers to be able to find "all and only" relevant legal cases, as well as their disposition (whether a law has been overturned or still stands).

In the realms of medicine and law it seems that strict adherence to controlled vocabularies is here to stay for generations to come. However, in many other disciplines we are seeing a paradigm shift. Users are tending not to take the time to look up subjects or descriptors and instead simply search the full text with tools like Google Web, Google Scholar, and Google Books.

Several observations should be made about controlled vocabularies and their use by various database vendors. The first observation is that there are great challenges when trying to decide whether to use a controlled vocabulary or not, and then which controlled vocabulary should be used. As we saw with the three social science disciplines of psychology, education, and sociology, there are sometimes great differences in nomenclature within these disciplines. What about contexts that are more universal? Library catalogs, for example, are collections of books and other materials across all disciplines: arts and humanities, social sciences, and science and engineering. Usually one controlled vocabulary is used in academic libraries to capture the "aboutness" of these works: the *Library of Congress Subject Headings*. But in using a generalized nomenclature set, subject-specific nuances are not captured. The important point in this observation is that controlled vocabularies are many, they vary in scope and applicability, and they are not always applied in every context.

The second observation we need to make is that many databases, through "smoke and mirrors," create the impression that they are using a principled thesaurus and applying it consistently throughout their product, but in fact this is not the case. This is not meant to criticize them, for they are doing the best they can with what they have to work with. But we need to be aware of what is really happening in aggregated databases like those produced by vendors like EBSCO, ProQuest, and Gale.

Unlike databases like PubMed and Westlaw, which get their records from a single stream that they control, aggregator content comes from many sources, some of which they have more control over than others. To make it appear that they have a semblance of vocabulary control, subjects are first captured out of individual index records, then are "back-generated" into a master index and "normalized" (brought into a uniform style) to the extent possible. Rather than individual index records being carefully examined (an impossible idea when you consider the scale of records vendors deal with), clusterings of like subjects are mashed together, and outliers are dealt with on an ad hoc basis. All three of these vendors have done a great job of cleaning up records that only a few years ago were a mess. My point here is that controlled vocabulary sometimes is superb, especially in fields of medicine and law, but many times it is overplayed, because millions of records are massaged by computer and only nominally overseen by human eyes.

SUBJECT HEADINGS VS. SUBJECT DESCRIPTORS

A century ago books were carefully cataloged and a small number of subjects were assigned per work. But for every subject assigned, another set of catalog records had to be typed out and filed into those handsome card catalog trays. For this reason the infamous "rule of three" was often applied, stipulating that no more than three subjects should be applied to a book. And the subjects were precoordinated headings. By precoordinated I mean that they had one or more semantic notions per subject heading.

We can see this in any library catalog. Searching for the subject cats, we see results like these:

Cats (one semantic notion)

Cats—Aging (two semantic notions)

Cats—Anatomy (two semantic notions)

Cats—Anatomy—Atlases (three semantic notions)

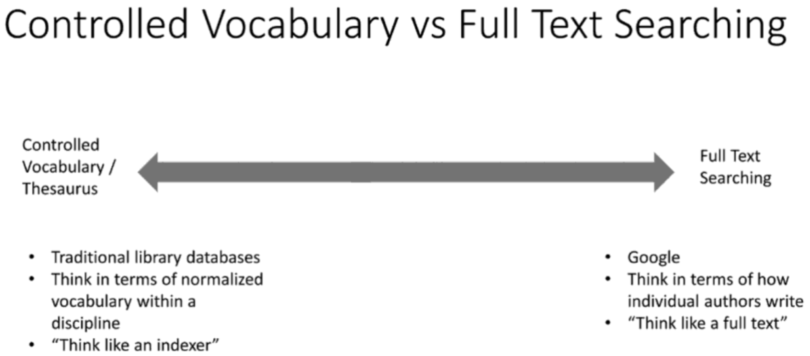
Cats—Anatomy—Juvenile Literature (three semantic notions)

Cats as Laboratory Animals—Law and Legislation—United States (three semantic notions)

Cats—Fractures—Treatment—Handbooks, manuals, etc. (four semantic notions)

In order to keep the subjects to three or fewer, it is necessary for catalogers to precoordinate the terms—that is, to combine notions together on a single line. In some cases six or more subjects can be precoordinated. Here is example of six semantic notions in a precoordinated Library of Congress subject heading: United States—Armed Forces—Yugoslavia—Pay, allowances, etc.—Taxation—Law and legislation. Yet even with all this effort to capture the "aboutness" in as few lines as possible, it is still failing to capture more granular subjects that are dealt with within books.

In contrast to the idea of subject headings are subject descriptors. Subject descriptors have one and only one semantic notion. In a descriptor world, the subjects are not precoordinated by a cataloger or an indexer. In the subject descriptor world the coordination must be done by the searcher. Thus, it is a postcoordination method. This means that if one were trying to search for books under the Library of Congress Subject Heading Cats—Fractures—Treatment—Handbooks, manuals, etc., searching would now need to be done



Controlled Vocabulary vs Full Text Searching

Controlled Vocabulary / Thesaurus

Full Text Searching

- Traditional library databases
- Think in terms of normalized vocabulary within a discipline
- "Think like an indexer"

- Google
- Think in terms of how individual authors write
- "Think like a full text"

FIGURE 1.2. Contrast in Thinking When Searching by Controlled Vocabulary vs. Searching Full Text.

It needs to be clearly stated that strategies for framing a subject descriptor search differ greatly from a full text search using Google. Under a subject descriptor method, searchers must think broadly about the topic. Consulting the thesaurus for the discipline, if there is one, is essential to understanding the normalized accepted nomenclature for the discipline. Searching should not be done by idiosyncratic terminologies used by individual authors, but by regularized terms accepted by searchers within the discipline.

On the other hand, full-text searching using Google is done at the other end of the spectrum. Searchers must place themselves, as much as possible, in the mind of the author, seeking idiosyncratic turns of the phrase and using a generous amount of synonyms. If one is searching for texts written in the 1920s, for example, terminology for racial groups and local customs will differ greatly. This is the difference between thinking like a "full text" and thinking like a cataloger or an indexer (Figure 1.2).

Library databases increasingly allow for full-text searching across content to which they provide full text. This proximity searching is quite powerful because users can control relevancy by stipulating how far keywords are from each other. For example, Term A within 20 words or 5 words of Term B, terms within the same sentence, or terms within the same paragraph. Whereas Boolean operators AND, OR, and NOT are nearly universal among online databases, proximity operators are not, and each vendor's product needs to be thoroughly studied in order to have confidence in proximity searching. In the world of Google, the library search operators do not work (not that users had really learned them anyway). Google's relevance ranking is supposedly smarter than any librarian. I'm not so certain of that, but that's the reality we have in Google's world.

much more than Google actually is able to crawl, is completely invisible to Google. Some have estimated that this hidden content is 500 times what is findable in Google (Bergman 2001). There are many reasons for this, and we need to discuss each of them. Failure to understand why Google does not find everything will only perpetuate the myth that everything and anything can be found with Google.

Google Is Polite

Google is not always wanted in all parts of the world. Perhaps you have seen news stories about China blocking all access to Google, favoring its own powerful search tool, Baidu. Lawsuits, both domestic and abroad, occasionally mandate that Google take down certain Web content because it is offensive, illegal, or disputed as part of legal actions. These kinds of actions have a big impact on Chinese students who return home after studying in other countries, or on researchers visiting China, but otherwise will have little impact on researchers in the United States.

Google will not crawl where it is not wanted. Google pays attention to robot exclusion protocols. A robot is another term for a Web crawler. One of the oldest of these protocols, still in existence today, is the robots.txt files posted on the root domains of many major Web sites. This file tells search engines where they should not crawl. Try this: go to your favorite Web site, and after the forward slash from the root directory, add robots.txt. For my university, the University of Denver, the root domain is: www.du.edu. Adding robots.txt to the root Web address gives us this URL: http://www.du.edu/robots.txt. Here you will see portions of the university Web site that Google need not bother to crawl, either because it is a waste of resources or because it doesn't add anything to the discoverability of information the university wants discovered. Here is what shows up on that page (Figure 2.1).

Not all Web sites employ this old technology. Many have newer tools to exclude Google and other search engines. But as a fun experiment, try to see how many robot exclusion files you can discover. Here are just a few examples (Table 2.1).

The point of this is that Google stays out when it is not wanted, accounting for at least a small reason why not all information is available to Google.

Technology Exclusion

Google is not able to go where the technology does not permit. Many databases are closed to search engines because of the way they work. Other databases contain information that, even if they could be crawled, would be meaningless. For example, the U.S. Naval Observatory has a database (http://aa.usno.navy.mil/data/docs/RS_OneYear.php) that gives sunrise/sunset

3

Searching Google Web

When I started teaching Internet Reference in 1999 Google was not the search engine I favored. In 1999 the World Wide Web was just six years old, and the popular search engines were AltaVista and soon after that AllTheWeb (known as Fast.com). Google existed at that time, but I paid little attention to it. But within a few years I became a convert: Google had figured out the relevance ranking magic and was continually developing new ideas through Google Labs. It was evident that Google was developing an interest in searching and discovery beyond what other search engines were interested in accomplishing. Labs was retired in 2011, but while it existed, it demonstrated the excitement and determination Google had in developing new ideas.

As time went by Google-like searching became so instantiated in culture that library search engines had to amend their search capabilities to keep up with user expectations. For example, the phrase contained within quotes, a search engine staple, is now a standard search feature in many, if not most, library-subscribed databases.

BASIC SEARCH TECHNIQUES

As computers become more sophisticated, formulation of search strategies evolved as well. In the early days of library searching users were taught the Boolean operators: AND, OR, and NOT. Along with that they were taught how to frame and use the best keywords to retrieve quality results. Now natural language searching changes things. Google is continually doing research to improve natural language searches (Google 2016). By natural language searching we mean just typing your question in the search engine box as you would as at the reference desk: "How many people live in Colorado?"; "What is the current consumer price index?"; or "How can I find contemporary reactions to Lincoln's Gettysburg Address?" These natural language searches actually work quite well in Google Web.

Although I don't recommend natural language searching for academic content, it can work some of the time. Most of the time it's best to search by noun forms and to stay away from other parts of speech like verbs, adverbs, adjectives, and prepositions. Some of the questions we pose in the academic world are replete with complexity—things like causation, correlation, implications, and other relationships. Perhaps future semantically based search engines will be able to sort out these complexities, but for the time being, it's best to stick with noun forms.

POWER SEARCH TECHNIQUES

For many users, power searching means going to the Google Advanced Search page. Google sometimes changes the way to get to this page, so the best way to get there is to search *google advanced search*. I find this page a bit painful to use. For this reason I think it best just to learn basic power search techniques and use them directly in searching. Here is a brief overview of the most useful power-searching techniques. These will be dealt with in greater detail with examples later.

Phrase Searching

Phrase searching is accomplished by enclosing your search term within quotes. When we say "quotes," we mean what is sometimes referred to as "double quotation marks." As you read this book you see smart, or curly, quotation marks used within the text. However, these must not be used in Google. Copying and pasting items containing curly quotes from Microsoft Word into a Google search box will sometimes give undesirable results. Although placement of a phrase in quotes is very powerful as a constraining mechanism, it should not be overused. Only enclose a phrase in quotes if it is really a "frozen phrase" in the language in which you are searching. "United States" is a frozen phrase, as we never say "the States that are United." However, "first amendment right" would not be a good idea to enclose in quotes, as the same idea might be framed as "first amendment gives us the right" or "the rights of the first amendment."

Site-Specific Searching

Most researchers I speak with don't realize that Google doesn't even let you see beyond the first 1,000 results, at best. Assuming that you retrieved 1 million results with your search, and assuming that you had many years to sift through results, Google prohibits you from looking at those results.

To test this, set your Google search results to 100 results per page, just to make this task easier. Now perform any broad, general search in Google, go to the bottom of the page, and you will see up to 10 pages to which you can navigate. The first page should have results 1 to 100, the second page results 101 to 200, and so on. Usually the results will stop far short of 1,000—maybe around 600 to 800 results. What this means is that if the results affecting academic research are not in these top 1,000 results, you will never see them. Throwing more words at Google will certainly bring different results to the top, but it may also keep out results that would have helped you. There must be a better way—and there is!

Site-specific searching means that you access Google's indexing of only a specific site. For example, it is often difficult to locate documents at my university, the University of Denver. But if I do a site-specific Google search using the syntax: *site:du.edu*, followed by my keywords, the result set is only results from the du.edu Internet domain. This syntax has a couple of rules that must be followed religiously: "site" must not be capitalized; and there must be no space after the colon. Technically it is okay to have a "dot" after the colon. For example, we can search *site:.gov* to find U.S. government information, but I never teach this as a best practice. When I teach in front of groups, I fear that the dot may be confused with a space, so I always omit it.

File Type Searching

Another way to escape Google's restriction on viewing 1,000 results or fewer is to limit by file type. File type or file format may be more familiar to those who use Windows-based computers. Although Macintosh and Windows computers have the option of concealing file type extensions, those who opt to see the extensions are familiar with these common three-letter extensions after file names: Adobe Acrobat (.pdf), Microsoft Word (.doc or .docx), Microsoft Excel (.xls or .xlsx), Microsoft PowerPoint (.ppt or .pptx), and generic text file (.txt). Limiting by these common file types with Google's syntax is most helpful in isolating research-related content. Very often substantive reports, studies, and articles are posted on the Web in Adobe Acrobat or .pdf format. Using syntax similar to site-specific searching, we can restrict results to .pdf format like this: *filetype:pdf*. The Microsoft file types mentioned earlier have two versions: the older version without the final "x," and the newer file types with the "x." Both of these need to be searched on their own.

It is also possible to search by less common file types such as WordPerfect (.wpd), Lotus AmiPro (.ami), generic rich text format (.rtf), Apple Keynote presentation (.key), and so forth. Table 3.1 summarizes these three most important power-searching strategies.

Other Considerations

There are those occasions when the advanced search interface is essential. For example, if searching for local information, such as news local to a

Index