Cross-Examining Google Scholar

***Xiaotian Chen** is an Electronic Services Librarian and Associate Professor at Bradley University in Peoria, Illinois. **Kevin O'Kelly** is Reference and Community Languages Librarian at the Somerville Public Library in Somerville, Massachusetts.*

*Correspondence concerning this column should be addressed to **Karen Antell**, Head of Reference & Outreach Services, and **Molly Strothmann**, Social & Behavioral Sciences Librarian, University of Oklahoma Libraries, 401 W. Brooks St., Norman, OK 73019; email: kantell@ou.edu and mstrothmann@ou.edu.*

If you ask a typical casual information user to name her first or favorite reference source, the answer is likely to be Google, and with good reason: Google's ease of use and ubiquity have opened up a world of information that formerly was trapped within book covers in libraries. However, when it comes to serious scholarship, can Google provide adequate access to research articles? Or do librarians still need to select specialized abstracting and indexing products and teach researchers how use them? In this installment of "Taking Issues," an academic librarian and a public librarian debate the strengths, weaknesses, and idiosyncrasies of Google Scholar.—*Editors*

Chen:

The topic of this column is "Is Google Scholar a reliable resource for scholarly research?" but it is not fair to discuss whether Google Scholar is a reliable resource for scholarly research without scrutinizing other resources. By singling out Google Scholar for scrutiny, we give other resources a free pass. The fact is, no resource is perfect, and every resource has weaknesses and errors. The salient question is, does Google Scholar have a higher percentage of errors or gaps than other resources? So far, we have no statistical data from empirical studies to prove that other resources have fewer gaps or errors than Google Scholar.

Researchers have already accepted Google Scholar and actually use it more often than most subscription-based abstracting and indexing (A&I) services. A survey by Hightower and Caldwell found that Google Scholar is among the three most popular A&I services that science researchers use (the other two being Web of Science and PubMed).[1] Various empirical studies have proved that Google Scholar can cover almost all the journals covered by subscription-based A&I services, especially for issues published during the past quarter century. Meier and Conkling's empirical study of 120 samples from Compendex, an engineering A&I resource, found that Google Scholar covered 90 percent of publications indexed by Compendex after 1990 and that the publications Google Scholar missed are mostly non-journal and non-English publications.[2] Walters compared Google Scholar with seven subscription-based A&I resources by searching 155 core articles on later-life migration published between 1990 and 2000. He found that Google Scholar retrieved the highest percentage of samples, 93 percent, outranking the best-performing subscription-based A&I (Social Sciences Citation Index) by a whopping twenty-seven percentage points.[3] In my own study of 400 samples, Google Scholar

was able to retrieve 98 percent to 100 percent of scholarly articles randomly generated from eight databases.[4] On the other hand, critics who assert that Google Scholar is not a reliable resource for scholarly research base their assessment primarily on anecdotal examples rather than statistical data. Here is my challenge to those who question Google Scholar's coverage of scholarly journals or its retrieval capability: name one scholarly journal that Google Scholar does not index, and name one article that Google Scholar cannot retrieve.

Many publications have exposed flaws in Google, Google Books, and Google Scholar, but subscription-based resources and services have not been subjected to similar scrutiny. Google Scholar is by no means perfect, but neither is any subscription-based resource. All A&I services, full-text databases, and library catalogs have errors, such as wrong dates, wrong page numbers, wrong volume and issue numbers, and even missing articles or issues. OpenURL link resolvers have been found to provide successful linking 80 percent of the time,[5] but nobody has characterized OpenURL linking as harshly as Google products have been described in the press and in the library literature: "Metadata mega mess in Google Scholar"; "Google's Book Search: a disaster for scholars"; "Google Scholar's ghost authors, lost authors, and other problems."[6] Why? Could it be that librarians hold the products that we select and provide for our users to a lower standard than we hold the competitor that involves no library participation?

Another reason that it may not make sense to claim that Google Scholar has more errors than other A&I products is that Google Scholar does not actually *create* bibliographic data as A&I services traditionally do. Rather, Google Scholar simply crawls and retrieves bibliographic records from journal websites (such as ScienceDirect and Wiley Online Library), aggregated packages (such as JSTOR and Project MUSE), free A&I resources (such as PubMed), institutional repositories, and some subscription-based A&I services that allow Google to crawl their content. Therefore, Google Scholar's "metadata mega mess" probably comes from publishers, vendors, and institutional repositories, and the "disaster" of errors some librarians see in Google probably comes from their local catalogs and WorldCat.

O'Kelly:

I don't think anyone wants to give other resources a "free pass," but I do think it's sensible to cast an especially critical eye at Google Scholar because many researchers use it simply due to its convenience as a Google product—not because they have critically assessed its value. As you note, many researchers use Google Scholar more often than they use subscription databases. These researchers deserve to know whether Google Scholar is up to the job. Certainly other A&I services are not infallible, but they have an incentive to avoid mistakes that Google doesn't: they exist to provide bibliographic information. Access to articles and article citations is the product they sell, but Google Scholar is just one of Google's myriad features. Frankly, in the case of Google Scholar—as with Gmail, Google+, and all other Google services—*we're* the product.

As for your assertion that Google is not responsible for its "metadata mega mess": no doubt some of the errors in Google Scholar are in the original bibliographic records retrieved by Google, but by no means all. As Peter Jacso of the University of Hawaii points out in *Library Journal*, Google Scholar's developers rejected the metadata offered by publishers and A&I companies. Instead, the records in Google Scholar are created by Google Scholar's parsers (software programs that analyze language), which have been known to create author names from menu options.[7] In one of the most infamous examples, "Please login" became an author name: "P. Login." Admittedly Google has corrected many "ghost author" mistakes, but to use Google Scholar is to run the risk of error propagation simply because its information has been copied from other sources. Why prefer second-hand information if first-hand is available?

You note that according to one study, many of the publications that slipped through the Google Scholar net were "non-English." I consider that a serious omission. Although English is the dominant language for academic publishing, access to materials in other languages is critical in some fields of study.

Chen:

With regard to content, Google Scholar probably focuses more on "scholarly" items than many A&I products do. The two empirical studies mentioned above found that Google Scholar covers basically all the journals in ERIC and Compendex but does not index some of the non-scholarly-journal items that these two databases include. ERIC and Compendex index many obscure items that are not even publications. For example, during a test search of ERIC, I retrieved a Pennsylvania county report on local school curricula; a Texas state legislature report; a medical school salary survey; a report on a community college's application, acceptance, and registration processes; a child care survey; and a congressional hearing on long-term care. ERIC and Compendex are by no means the only databases that include obscure items. EconLit indexes items called "working papers." EBSCO's Academic Search Premier includes summaries of reports. CINAHL indexes discussion lists and some other nonstandard publication types. These non-scholarly-journal items or non-publication items may help inflate the contents of A&I products and may even be useful to researchers, but modern A&I services seem to have started to provide the kind of broad coverage for which general search engines used to receive criticism. On the other hand, Google Scholar has apparently become more selective and better focused on "scholarly" items.

One major advantage of Google Scholar compared with traditional A&I products is that, in the age of open access, Google Scholar is far better at covering institutional repositories—not only in terms of identifying content, but more importantly, in facilitating access to free full text. Google Scholar offers free full text availability indicators when it retrieves items from institutional repositories. As more and more institutional repositories are being launched and more open access articles are becoming

available on the Internet, this advantage of Google Scholar will become more useful to researchers.

Another Google Scholar advantage is years of coverage. Most traditional A&I services have a starting year limit, excluding from their databases anything published before that year. (PsycINFO, Web of Science, and Chemical Abstracts are among the A&I products with the longest coverage periods, with starting years of 1887, 1900, and 1907, respectively.) Google Scholar does not have that limitation. It crawls publishers' websites, databases, and institutional repositories, and it retrieves items that they contain regardless of the year of publication. As most major publishers have begun posting tables of contents of their earliest journal volumes online, Google Scholar is able to discover those sources. For example, the first year of *European Journal of Organic Chemistry*, 1832, is available on Wiley Online Library. Google Scholar can retrieve articles published in this journal in its starting year: no other A&I can do that.

I cannot agree with you on the mission differences of subscription-based A&I resources and Google Scholar. You note that commercial A&I products exist to provide bibliographic information and that access to articles and article citations is the product they sell. Actually, Google Scholar's mission is exactly the same as that of subscription-based A&I services. The difference lies in their indexing methods. Subscription-based A&I products still use a traditional indexing model: they collect bibliographic data and store it in closed storage. Google Scholar, on the other hand, crawls the open Internet and targets the servers of publishers, aggregators, database owners, institutional repositories, academic and research websites, and so forth. In addition, you also assert that subscription-based A&I products have incentives to avoid mistakes that Google doesn't, but many A&I vendors do not own A&I content and therefore are powerless to correct errors. For instance, EBSCO, OCLC, Ovid, ProQuest, and other A&I vendors do not own MLA International Bibliography, so they have neither the incentive nor the right to correct any errors in MLA's indexing. In fact, almost all electronic resource license agreements state that vendors are not responsible for the content providers' errors.

Finally, we cannot compare products without comparing prices. As the methods of knowledge dissemination evolve, the unique value of traditional A&I products is decreasing because almost all publishers post their journals' tables of contents on the Internet, and Google can index them. Yet the prices of traditional A&I services continue to rise, even though technology has made indexing work easier. Therefore, the time is ripe for libraries to shift from dependence on commercial A&I vendors to reliance on free A&I services, including but not limited to Google Scholar. As the Hightower and Caldwell survey showed, Google Scholar is now one of the top three A&I choices of science researchers; this is a trend that librarians cannot halt. Moreover, this shift also will allow libraries to conserve financial resources for information resources and services (such as full text resources) that do not have viable free alternatives.

O'Kelly:

Frankly, the breadth of coverage provided by ERIC and Compendex sound like advantages rather than defects: as you note, some of those items "may even be useful." If a scholar wants to research (as an example) funding of Texas schools, a report from the Texas state legislature may well be essential. With regard to your other points, everything you say about Google Scholar in regard to years of coverage and price is true—and desirable from the points of view of both institutions and researchers—but I don't think it's time just yet to kick out EBSCOhost and Gale. Google Scholar is still an immature search tool, in my opinion, whereas EBSCOhost and other database companies have years of experience and employ trained people, not algorithms, to organize information and enable retrieval using controlled vocabularies. As Jerry Gray of the University of Indiana puts it,

> Content found through indexing services, in academic repositories, and on government web sites undergoes some level of scrutiny, either by peer review, editorial process, or selection standards, ensuring a level of accountability for the quality of what is presented. However, with the enduring popularity of search engines like Google Scholar that do not conform to a validation process, the line between scholarly science literature and pseudoscience is no longer so clear.[8]

Quite bluntly, we don't know how Google does what it does, and even the people working at Google don't seem to know entirely what they are doing. Before it was revised in April 2012, the text on Google Scholar's Metrics page read, "since Google Scholar indexes articles from a large number of websites, we cannot always tell where (or if!) a particular article has been published." The deletion of this disclaimer shouldn't lead us to assume that Google Scholar employees have rectified the problem.[9]

Finally, there is the issue of how Google Scholar orders search results. The more often a paper or article has been cited by other researchers, the higher its page rank in the search results. This is problematic for two reasons. First, it creates a "Matthew Effect"[10] for research: work that is already influential becomes even more widely known by virtue of being the first hit from a Google Scholar search, whereas possibly meritorious but obscure academic work is buried at the bottom. As a thought experiment, imagine an alternate universe in which Google Scholar existed in the late seventeenth century. A student searching for basic information on physics, such as the laws of motion, would retrieve works by and about Aristotle because up until the Renaissance, most scientific questions were settled not by experimentation but by asking, "What did Aristotle say about it?" Meanwhile, the work of a Cambridge scholar named Isaac Newton probably wouldn't even appear on the first page of results.

Google Scholar's system of basing page rank on the number of citations is problematic for a second reason: it's quite easy to game the system. Recently, a group of Spanish

researchers created six documents by a non-existent author, citing a (real) academic's publications. The resulting snowball effect added 774 citations to the cited scholar's work.[11]

Although I agree with many people that Google Scholar (like Wikipedia) can be useful for a preliminary search, for solid research I'll stick with my library's databases.

References

1. Christy Hightower and Christy Caldwell, "Shifting Sands: Science Researchers on Google Scholar, Web of Science, and PubMed, with Implications for Library Collections Budgets," *Issues in Science and Technology Librarianship* 63 (2010), www.istl.org/10-fall/refereed3.html.
2. John Meier and Thomas Conkling, "Google Scholar's Coverage of the Engineering Literature: An Empirical Study," *Journal of Academic Librarianship* 34, no. 3 (2008): 196–201.
3. William Waters, "Google Scholar Coverage of a Multidisciplinary Field," *Information Processing & Management* 43 (2007): 1121–32.
4. Xiaotian Chen, "Google Scholar's Dramatic Coverage Improvement Five Years after Debut," *Serials Review* 36, no. 4 (2010): 221–26.
5. Cindi Trainor and Jason Price, "Chapter 3: Digging into the Data: Exposing the Causes of Resolver Failure," *Library Technology Reports* 46, no. 7 (2010): 15–26.
6. Péter Jacsó, "Metadata Mega Mess in Google Scholar," *Online Information Review* 34, no. 1 (2010): 175–91; Geoffrey Nunberg, "Google's Book Search: A Disaster for Scholars," *Chronicle Review* 31 (August 2009), http://chronicle.com/article/Googles-Book-Search-A/48245 (accessed Feb. 7, 2013); Peter Jasco, "Newswire Analysis: Google Scholar's Ghost Authors, Lost Authors, and Other Problems. Why the Popular Tool Can't Be Used to Analyze the Publishing Performance and Impact of Researchers," *Library Journal* online, September 24, 2009, www.libraryjournal.com/article/CA6698580.html.
7. Jasco, "Newswire Analysis."
8. Jerry E. Gray, Michelle C. Hamilton, Alexandra Hauser, et al., "Scholarish: Google Scholar and its Value to the Sciences," *Issues in Science & Technology Librarianship* no. 70 (Summer 2012), www.istl.org/12-summer/article1.html.
9. Ibid.
10. "For whosoever hath, to him shall be given, and he shall have more abundance: but whosoever hath not, from him shall be taken away even that he hath." Matthew 13:12.
11. Emilio Delgado López-Cózar, Nicolás Robinson-García, and Daniel Torres-Salinas, "Manipulating Google Scholar Citations and Google Scholar Metrics: Simple, Easy and Tempting," *EC3 Working Papers* 6 (May 29, 2012), arXiv:1212.0638, http://arxiv.org/abs/1212.0638.