Chapter Twenty-Three

Does Google Scholar Help or Hurt Institutional Repositories?

Peace Ossom Williamson and Rafia Mirza

Librarians often act as though Google Scholar is our new frenemy; however, the reality is somewhat more complicated. While it is true that Google Scholar's requirements are not always transparent to us (us being libraries, archives, institutional repositories), enough research has been conducted that we can now make an educated guess as to how to organize the content in institutional repositories (IRs) and archives in ways that Google Scholar can index more effectively. Google Scholar offers many options for libraries to make their IR collections discoverable. It is true that there has been a history of Google Scholar lacking in transparency and structure; what's more, Google Scholar citations profiles are largely accomplishing the same objectives as IRs. This chapter looks at the effect these Google programs have on IRs and how libraries can respond.

Google Scholar has gained its powerful position because libraries have yet to create an effective means of searching across IRs. The combination between this lack of otherwise-created searching capability and the large general public preference toward Google resulted in Google Scholar becoming the default search mechanism for scholarly works across a broad range of locations. Librarians' learning to navigate in Google's territory arises from the growing movement toward open access content, especially in scholarly work. This open access movement arose from the ongoing serials and access crisis, where the costs of serials are increasing rapidly. Most libraries' budgets cannot accommodate these cost increases, and they must cut access to valuable scholarly works. Now, only the more affluent—typically Western—institutions can afford the most expensive resources. As scholars continue to

write for impact rather than recompense, publishers profit handsomely from scholarly works.

Open access differs from this traditional structure, as scholarly works are provided without charge to readers, lessening the purchasing pressure on libraries. Publishers are slow and resistant to provide open access to their publications, and, in response, libraries are storing the works of their affiliated scholars in repositories in order to provide ease of access and archiving of these works. These repositories are institutional, as the resources present are works of individuals from that particular institution. IRs contain the metadata and full text of scholarly work produced by people affiliated with that institution. Works that are otherwise hidden behind a paywall, like articles in medical journals, or works that are largely inaccessible to the greater public, like theses and dissertations, are made available in order to increase the impact of work stemming from a particular institution. IRs include a wider range of resources than traditional journals or publications. Student work, gray literature, and academic and professional presentations are also included.

While there are major differences in the more than three thousand repositories existing across the globe, IRs provide research to the greater community through the green open access model—that is, there is neither a cost to deposit works nor a cost to access works. Green open access puts control of publication in the hands of the research community and can improve dissemination, if done well. This increases the number of citations, as these resources are no longer limited to users affiliated with Western institutions with greater amounts of financial capital. In response, publishers are beginning to offer gold open access models. Gold open access models are models in which authors may pay some of the costs of publication, but access to full text within these publications have no cost. While somewhat beneficial, IRs have a greater benefit in that they cut publishers' ability to control pricing, return control of research to the hands of the researcher, and improve resource "sustainability and stewardship" through the management and preservation of original output (Poynder 2014).

INDEXING AND DISCOVERY

Google consistently has more than 65 percent of the share in search engine use, and this rate has also been true of Google Scholar (Arlitsch and O'Brien 2012). Because of Google Scholar's dominance, many users are able to find useful literature from IRs without any prior knowledge of the purpose or existence of these repositories. It also provides about three-fourths of traffic to IRs (Poynder 2014). Furthermore, it provides a proven location for centralized searching and a familiar interface for most users. The indexing of IRs by Google Scholar is of importance due to their relationship. Google Scholar

is a search engine independent of Google Search in that its purpose is finding scholarly works regardless of where they are located on the web. In that pursuit, "Google privileges information from reputable sites, such as those with '.gov' or '.edu' or 'ac.uk' domain names" (Dawson and Hamilton 2006). Google Scholar has been heralded in its ability to crawl publisher pages, open access databases, IRs and other repositories, and author websites in addition to other locations for research literature. Although designed specifically for simple searching, it is an outstanding resource for obtaining gray literature (Giustini and Kamel Boulos 2013). In fact, since Google Scholar shows the primary version of the article (which is generally not the preprint in the IR, but the post-print at the publisher's site), it may be better for IRs to focus on gray literature (Arlitsch and O'Brien 2012).

Because Google Scholar is the most preferred method of searching for scholarly works and, therefore, the primary indexing service (even if by default), IR administrators must be aware of its indexing methods. In order to increase discoverability, IRs need to take into account what information Google Scholar is using for indexing and how it interprets our metadata. If we do what it wants and accommodate its needs, it is great for our IRs. If we do not, then it can greatly diminish our discoverability by making our IRs almost completely invisible. The major setback is that librarians have to guess how this is done. Google is secretive about its algorithm, minimizing users' ability to perform expert searches. It has advanced searching capabilities; however, it has very few limiters and no controlled vocabulary. Nevertheless, Dawson and Hamilton (2006) point out that some sites, such as Physics Finder, have optimized their pages to get higher page rankings, while keeping in accordance with their metadata schema.

When it comes to the public interface, it is also difficult for users to determine which materials fall within or outside of Google Scholar's scope. Google and Google Scholar indexes are separately created and maintained due to Google Scholar's focus on peer-reviewed articles, books, white papers, patents, and legal reports; therefore, results vary vastly when the same search is completed within Google or within Google Scholar. Results are organized by key words, the number of citations, and the number of previous clicks on the links; therefore, many users struggle to retrieve newly published materials or relevant research. Searching for titles or author names is a much more successful method of using Google Scholar, while topic and keyword searches remain largely unpredictable. There are ways to increase the rankings of articles in repositories by including information such as subject headings.

OPTIMIZING DISCOVERABILITY

The indexing ratio for an IR in Google Scholar is the sum of the IR's unique URLs in Google Scholar over the total sum of unique URLs in the repository. This ratio is typically low in the average IR due to the conflict between Google Scholar metadata requirements and the procedures of many IRs. Search engines are severely restricted to searching text, and they cannot read text present within multimedia, JavaScript, or images; moreover, IR databases and servers must allow search engine crawlers to be present in the first place. The crawlers then follow links to the metadata, which is evaluated by Google Scholar algorithms, and these algorithms determine whether that information is added to the Google Scholar index (Arlitsch and O'Brien 2012, 64). Google Scholar provides guidelines for search engine optimization (SEO) on its site through improving the success of its crawling and indexing websites and repositories. Some of these guidelines recommend using up-to-date software and providing chronological lists of works and permanent links (Google Scholar 2010). Most notably, Google Scholar (2010) discourages the use of Dublin Core as a metadata schema. Arlitsch and O'Brien found that IRs that adhered to the indexing guidelines put forth by Google Scholar had an indexing ratio of 88 percent to 98 percent, while IRs that did not had a much lower indexing ratio of 38 percent to 48 percent (2012, 70).

The question then, of course, is, should librarians adopt wholesale a metadata schema that is not created by librarians? Should SEO for Google Scholar be a primary concern of ours? The short answer to these questions is a resounding, "Yes!" To a certain extent, we need to accept *satisficing* as the primary searching method of the average user; therefore, we cannot dictate how users come to our online repositories. The goal of an IR is to make the content more accessible; the goal is not necessarily prominence or recognition for the IR itself. In addition, Dublin Core works poorly as a schema for articles because it is open to inconsistent interpretation in practice. For example, Dublin Core includes one field for multiple citation data, including journal name and volume number, and there are no fields distinguishing document type (Arlitsch and O'Brien 2012). However, as is often said, "Metadata is a love note to the future," and we want to make sure IRs include the metadata that future librarians and archivists may find useful in addition to adapting to the ways our current users find resources in our IRs. We can do both by using current standards, such as Dublin Core, and by also adding whatever code or script overlays are needed for the materials in our IRs to be indexed by Google Scholar. We want to do more than merely pay lip service to interoperability, yet we do not want to go overboard in allowing Google Scholar to completely dictate our choice of metadata schema, as there is no guarantee that Google will always be the Internet's search engine of choice.

We want to prioritize future searchers as much as we prioritize current searchers.

ADAPTING TO METADATA SCHEMAS

Libraries must adapt to the online environment by updating access and assessment of resources in the IR. For example, Google Scholar gives users a direct link to the PDF full text of resources found, thereby bypassing any context. This circumvents the libraries' mechanisms used in keeping visitation statistics for PDFs that are separate from the HTML display. In response, libraries should add a PHP script to more accurately track usage statistics (Arlitsch and O'Brien 2012). Libraries can also use schemas—for example, PRISM or Highwire Press—that are better able to accept citation information. Furthermore, it is of utmost importance that IR administrators avoid errors including dead links or slow servers because when crawlers encounter errors on a site, they are less likely to return. Google Webmaster Tools in combination with assessment tools can reduce instances of errors and improve analysis of the impact of an IR on its institution.

Dawson and Hamilton use the term *data shoogling* to "refer to the process of rejigging, or republishing, existing digital collections, and their associated metadata, for the specific purpose of making them more easily retrievable via Google" (2006, 313). The four standards they espouse are:

1. Search engine optimization
2. Metadata cleaning
3. Metadata optimization
4. Metadata exporting

Basically, they argue that we should engage in consistent metadata, keeping in mind the constraints of the web page and what we do know about what web crawlers look for. This seems a reasonable argument. Dawson and Hamilton discuss the ways in which libraries can export their collection's metadata to a series of static pages (that get updated to reflect the collection if it changes) that get indexed in Google Scholar, increasing the likelihood that people will find the collections using subject terms. They give the example of the Glasgow Digital Libraries, whose collections received high ranking in Google Scholar without any external links, seemingly simply by having "search-engine-friendly design" (Dawson and Hamilton 2006, 319). If librarians include Library of Congress subject headings (LCSH) in the pages for their collections, it aids searchers because those LCSHs are phrases: if a searcher uses even part of that LCSH phrase in their search, they are more likely to find relevant content. Indeed, in the example of the Glasgow Digital

Libraries, users were finding content by searching generally on a topic, not by specifically searching for the holdings of the Glasgow Digital Libraries. Arlitsch and O'Brien (2012) found that if you expressed IRs' Dublin Core metadata schemas in HTML meta tags, those IRs had higher indexing rates. Thinking of search engines as "users with substantial restraints" in regard to what types of content they can view (multimedia, JavaScript, etc.) may be the best way for librarians to think about these indexing services.

Google Scholar is how many people find open access academic resources. As librarians increase the promotion of open access content, we should think about ways in which we can maintain metadata standards, but also be receptive to the ways in which search engine crawlers harvest sites such as IRs. While Google Scholar does not have a metadata HTML tag exclusively for LCSHs, if librarians include LCSHs in their HTML meta tags or titles, they will improve their sites' rankings.

ATTRACTING ORIGINAL CONTENT

In addition to discoverability, librarians must also be aware of the robustness of their IR. Many authors have yet to be convinced of the benefits of depositing their works in IRs. This is a difficult issue to overcome, which is why IRs consist of few primary articles. It becomes a burden with no benefit in situations where academics are already publishing in open access journals.

Low percentages of primary articles have a great effect on IR discoverability because IRs that provide a larger number of primary articles, especially gray literature, are more likely to be included in Google Scholar results and rank higher in these lists, thereby improving their indexing ratio (Arlitsch and O'Brien 2012). As indexing ratios—a measure of IR discoverability—increases, scholars are more likely to deposit their original work or primary articles. Attempting to change one or both of these two codependent characteristics can create a difficult cycle for libraries. In order for libraries to break out of this cycle, they must first convince academics to deposit their already-published works in the IR then convince those same academics that, from that point forward, they need to negotiate with publishers after the peer-review process for rights to the final edited copy of their works. Methods of convincing students and faculty members to deposit their works in an IR include the following benefits:

- An increase in their works' citation counts and potential impact
- The retention and preservation of published works
- The author's access to his own works, even if the institution's subscription is canceled

- The retention of the author's copyrights and use of one's own work in teaching and in self-promotion
- The reduction in plagiarism, since the original source is openly available for referral

A common response from academics is that they are already using the Google Scholar citations profile or a similar profile via other websites, such as Academia.edu . Google Scholar profiles were made available in 2011, and users are able to create a profile by selecting their publications in Google Scholar. The individual's publications are subsequently listed, and that user is provided with citation metrics and graphs.

It is difficult to know what citations are included in these indexes because the list of resources used in Google Scholar is privately kept. Google Scholar citation information can, in ways, be more comprehensive than traditional databases: Nicola J. Cecchino (2010) found results within Google Scholar that were not present in Web of Science. Conversely, the weakness that has caused a somewhat even playing field between Google Scholar citation profiles and IRs is the lack of authors who have opted into either system. While Google Scholar benefits from already having the metadata of many publications, it too must appeal to scholars in order for them to create the profiles that link these publications. It is likely much easier for a user to sign up for a profile with Google Scholar than it is for them to deposit their works and have a profile created for them in their IR, so it is crucial that we actively promote IRs across our institutions.

Advantages that librarians can share with academics who have or are considering Google Scholar profiles instead of depositing their works in the IR are as follows:

- IRs' clearly defined metrics of assessment in comparison with a Google Scholar citations profile, where Google does not share their methods of calculations and assessment
- Open access to most, and likely all, published works deposited in the IR
- Ability to deposit presentations and other scholarly works not included in Google Scholar results and analysis
- IRs' allowance for adaptation: when Google Scholar is no longer the preferred search method, IRs can adjust to requirements for discoverability on another site

Libraries that are able to succeed in early adoption efforts will have the greatest discoverability within Google Scholar and the greatest chance of growth in the future. Once Google Scholar is able to amass a large number of profiles and if academics continue the trend toward publishing open access

works, it will severely limit the likelihood of IRs ever being as robust as they need to be in order to get top spots in search-result lists.

ADVOCATING FOR LIBRARIES

Finally, libraries must continue to advocate for themselves. At one time, there was a push for Google Scholar to mark whether results were open access or not. This tagging would have allowed users to limit their results to open access content, largely increasing the visibility of IRs; however, Google did not choose to make that change. This may change in the future; after all, both Google and Google Image Search did not originally allow you to search by license, but those filters are now available in their advanced searches. Hopefully, in time, that filter will be added to Google Scholar.

In the meantime, mandates like the Office of Science and Technology Policy (OSTP) Increasing Access to the Results of Federally Funded Scientific Research memorandum ensure that open access initiatives will only become more and more important. The SHared Access Research Ecosystem (SHARE) resulted from universities and their libraries laying the groundwork to become more cohesive and structured in their use of IRs as they seek to grant public access to their universities' publicly funded output. This groundwork is based on the presupposition of the importance of IRs in this open access (OA) ecosystem and the importance of standardizing metadata requirements and ensuring those requirements are exposed to search engines and discovery tools (Association of American Universities et al. 2013). Libraries can ensure open access by increasing the robustness and infrastructure of their IRs and helping scholars gain control of their scholarships' dissemination, or they can leave it to others, such as publishers (with their proposed alternative response to the OSTP memo called CHORUS [CHOR, Inc. 2014]), which may not have the same focus on metadata and long-term access that universities and their libraries have.

While Google Scholar may not be completely transparent in the ways in which it indexes IR content, we have enough knowledge to:

- Get content by engaging in outreach to promote IRs for open access
- Focus on unique content (gray literature) or final versions of articles
- Make sure content is findable by optimizing our discoverability
    - ensuring indexing in Google Scholar through the use of metadata
    - shoogling content if necessary

CONCLUSION

In this chapter, we focused on what individual IRs can do to increase their indexing in Google Scholar, but we need to make sure we also emphasize initiatives such as SHARE that allow for the national library community to embrace standardized metadata for the open web and increase the profile of IRs and promote them as integral to the research process. Librarians can and should move toward a future in which we maintain our professional values of transparency and open access by staying aware of the ways in which changing technology impacts our efforts, as well as the shift of scholarly publications toward open access.

REFERENCES

Arlitsch, Kenning, and Patrick S. O'Brien. 2012. "Invisible Institutional Repositories: Addressing the Low Indexing Ratios of IRs in Google Scholar." *Library Hi Tech* 30 (1): 60–81.

Association of American Universities, Association of Public and Land-Grant Universities, and Association of Research Libraries. 2013. *SHared Access Research Ecosystem (SHARE): Development Draft*. June 7. http://www.arl.org/storage/documents/publications/share-proposal-07june13.pdf .

Cecchino, Nicola J. 2010. "Google Scholar." *Journal of the Medical Library Association* 98 (4): 320–21. doi:10.3163/1536-5050.98.4.016.

CHOR, Inc. 2014. "About CHORUS." *CHORUS: Advancing Public Access to Research*. http://www.chorusaccess.org/about/about-chorus/ .

Dawson, Alan, and Val Hamilton. 2006. "Optimising Metadata to Make High-Value Content More Accessible to Google Users." *Journal of Documentation* 62 (3): 307–27.

Giustini, Dean, and Maged N. Kamel Boulos. 2013. "Google Scholar Is Not Enough to Be Used Alone for Systematic Reviews." *Online Journal of Public Health Informatics* 5 (2): 214–22. doi:10.5210/ojphi.v5i2.4623.

Google Scholar. 2010. "Inclusion Guidelines for Webmasters." http://scholar.google.com/intl/en/scholar/inclusion.html .

Poynder, Richard. 2014. "Interview with Kathleen Shearer, Executive Director of the Confederation of Open Access Repositories." *Open and Shut?* (blog). http://poynder.blogspot.com/2014/05/interview-with-kathleen-shearer.html .