Standing on the shoulders of the Google giant: Sustainable discovery and Google Scholar's comprehensive coverage.

11/19/2015

The 11th anniversary of Google Scholar passed yesterday. **Max Kemman** *provides an overview of the growth and impact of the platform and also looks at why Google Scholar is virtually unrivaled. The scholarly community might ask whether it is entirely desirable that Google plays such an important role in the scholarly workflow. Not only does Google Scholar have a known effect on discovery and citation of articles, it could well be shaping academic writing and evaluation.*

"Standing on the shoulders of giants" has long been the metaphor of choice to describe the scholarly workflow of discovering, reading, and citing literature. However, for the past decade this workflow has been influenced significantly by the availability of academic search engines. In this field, the search giant Google has come out as the discovery mechanism of choice. How does "standing on the shoulders of the Google giant" impact the scholarly workflow?

Recent surveys have inquired about the use of discovery systems for academic literature. In a survey we conducted in 2012 with 288 Dutch respondents in the Humanities and Social Sciences, 88% used Google Scholar to some extent. JSTOR was a close second, used by 85%, although it was used less often. In a recent survey on 101 Innovations in Scholarly Communication, a preliminary finding from the first 1,000 responses was that 92% used Google Scholar. Here Web of Science was a distant second, used by 47% of respondents. Although such surveys do not necessarily show *how* Google Scholar is used, it is clear that a large portion of scholars searches for literature at Google Scholar. One research report concluded in 2013 that "library and publisher platforms were not central to discovery, but Google and Google Scholar were"*.*

Impact on the scholarly workflow

This shift from the library to Google surely must impact the results of the scholarly workflow. HighWire recently summarized this in the following points:

- Search is the new browse
- Full text indexing of current articles plus significant backfiles joined with relevance ranking to change how we looked and what we did.
- "Articles stand on their own merit"
- "Bring all researchers to the frontier"
- "So much more you can actually read"

One study found that all this resulted in doctoral students citing more literature since 2004, calling this ' the Google effect'. The Google Scholar team themselves found that with the availability of academic search engines, the impact of non-elite journals has grown, as well as the impact of older articles. Both results can partially be explained by the way Google Scholar presents literature: articles are represented by their titles and search snippets, putting less emphasis on the journal in which it is published. Moreover, Google Scholar ranks articles with more citations higher, meaning older articles have an advantage over newer articles. One study concluded that by doing so Scholar introduces a Matthew effect in the impact of older articles. Another notable choice for ranking this study found is that Scholar puts a high weight on words occurring in the article's title.

Image credit: Tony Hirst Google scholar results (CC BY 2.0)

Other changes in the scholar's workflow summarized by HighWire include a growth in the number of articles clicked on, especially to read abstracts, as well as a growth in diversity of areas clicked on. One interesting consequence of these two changes is that scholars might want to write more accessible abstracts for the wider audience that finds their article through keyword searches and who might be interested despite not being an expert in the author's field. In short, not only does Google Scholar have a known effect on discovery and citation of articles, it could have an unknown effect on the writing by authors as articles are increasingly ranked and evaluated on their titles and abstracts first.

Sustainable search

As is usually the case when scholars depend on an external entity for a very important task, the sustainability of Google Scholar has long been a worry. When Google Scholar lagged behind the new Google-logo design, scholars expressed concern on Twitter. The co-founder of Google Scholar, Anurag Acharya, has made it a case however that Scholar is really not in danger. For Google, Scholar is a relatively easy search problem, with a small user base, so maintaining the academic search engine is a small cost. Moreover, since many Googlers are ex-academics, Scholar gets a lot of sympathy from within the company.

Maybe the concern over Google Scholar's sustainability is thus not needed. Still, we might ask whether it is entirely desirable that Google plays such an important role in the scholarly workflow, and as such in science in general. A question remains over why it is so difficult to replace Google Scholar with an alternative. Numerous features can be mentioned to compare between Google Scholar and its competitors, but one feature with which Scholar stands out is comprehensiveness. Although it is unclear how much is in Scholar, it is clear that is has the largest coverage of all the available discovery systems. There are three reasons for this:

1. First, Google Scholar is essentially built on top of Google, meaning it is not limited to specific databases, but can work with a heuristic to decide whether or not to include something in Scholar: if it cites other academic work, and if other academic work cites it, it probably is academic work.

2. Second, because of this, in contrast to search engines offered by publishers, Scholar indexes works that are available on the web in open access (or possibly illegal) form. This means Scholar usually is able to offer the author's version of a paper.

3. Finally, because Google is such a giant, it has managed to get publishers to agree to be crawled, so it can also index the full-text of the publisher's version. This final point is what makes it difficult for scholars themselves to come up with a powerful alternative. For example, the recently launched Semantic Scholar looks interesting, but is limited to publicly available online articles. This means that such an undertaking misses a vast amount of literature, and is thus from the start already less attractive, whether its functionality is better than Google's or not.

If we as scholars are genuinely concerned with Google Scholar's sustainability, as well as with Google's dominance in the scholarly workflow, it seems to me the only solution is to push for open access availability not only of new articles but also of old ones. Only once a comprehensive open access database of academic literature can be developed, can we really open up the space for competition with Google Scholar. Although Academia.edu allows (some) search engines to crawl their website after a request for permission, there is a debate whether this would be the best approach for open access. Whether this will lead to better offerings of course remains to be seen, as comprehensiveness is not the only aspect of interest. But I would be very interested to see what we can come up with as discovery systems once the data is available. Maybe this is an unreachable dream, but if there is one thing we can learn from Google Scholar, it is that the publishers' monopoly over access to academic literature can be disrupted.

Featured image credit: Eduardo Paolozzi's *Newton, after William Blake* (1995), outside the British Library by John McCullough – Wikimedia, CC BY 3.0

Note: This article gives the views of the author, and not the position of the LSE Impact blog, nor of the London School of Economics. Please review our Comments Policy if you have any concerns on posting a comment below.

About the Author

Max Kemman *is a PhD Candidate at the University of Luxembourg. His research concerns the scholarly use of digital technology, focusing on how the research practices of historians are influenced by the interaction with digital technology and technologists. To follow his research see his blog at www.maxkemman.nl and his Twitter @MaxKemman.*