

Tracey Hughes interviews Anurag Acharya, lead engineer on Google Scholar



When I interned at Google last summer after getting my MSI degree, I worked on projects for the Book Search and Google Scholar teams. I didn't know it at the time, but in completing my research over the course of the summer, I would become the resident expert on how universities were approaching Google Scholar as a research tool and how they were implementing Scholar on their library websites. Now working at an academic library, I seized a recent opportunity to sit down with Anurag Acharya, Google Scholar's founding engineer, to delve a little deeper into how Scholar features are developed and prioritized, what Scholar's scope and aims are, and where the product is headed.

TH: Can you tell us something about how Google Scholar came about?

AA: Alex Verstak and I used to work on building Google's web index. This was very hectic work and after several years of it, the two of us took a break -- a sabbatical of sorts -- for a few months. Google Scholar came out of that sabbatical. I had already been working on including scholarly literature in Google's index. For the sabbatical, we worked on improving indexing, automatically extracting metadata and ranking for scholarly literature. Our hope was to weave this information into Google web search. But "there's many a slip betwixt the cup and the lip." A working demo we sent out internally became popular and Google Scholar was born.

TH: What is your vision for Google Scholar?

AA: I have a simple goal -- or, rather, a simple-to-state goal. I would like Google Scholar to be a place that you can go to find all scholarly literature -- across all areas, all languages, all the way back in time. Of course, this is easy to say and not quite as easy to achieve. I believe it is crucial for researchers everywhere to be able to find research done anywhere. As Vannevar Bush said in his prescient essay "As We May Think" (*The Atlantic Monthly*, July 1945), "Mendel's concept of the laws of genetics was lost to the world for a generation because his publication did not reach the few who were capable of grasping and extending it; and this sort of catastrophe is undoubtedly being repeated all about us, as truly significant attainments become lost in the mass of the inconsequential."

TH: What disciplines does Google Scholar cover?

AA: We believe Google Scholar covers all major disciplines. To get a flavor of this, try restricting your search to one or more broad areas from the Google Scholar [advanced search page](#).

TH: Do you index meeting abstracts?

AA: Yes we do. In some research areas, including my own, conference proceedings have become the primary means of communication. We include conference abstracts and articles, journal articles, preprints, dissertations, books and so on.

TH: How does Google Scholar rank search results?

AA: We try to rank search results as a researcher in the area would. We take into account many factors, including who wrote the article, where it was published, how relevant the article is to the query, and what other articles have said about it.

TH: What does "cited by" mean? How do you come up with the "cited by" number?

AA: The "Cited by" link allows you to see a list of other articles which have referenced the work. Each Google Scholar result presents a body of work. Each body of work may have many manifestations or descriptions. For example, it could be initially described in a conference, then as a preprint, then as a journal and finally as a part

of an anthology. We try to group all such manifestations. We then figure out which works have referenced which other works. The "Cited by" number is the number of such references. Note that since we group multiple manifestations of a work, a "Cited by" list can and often does include references to different manifestations (e.g. a preprint and a journal article).

TH: Does Google Scholar work with bibliographic management software?

AA: Yes, it does. The Google Scholar [preferences page](#) allows you to select your bibliographic management software (we currently support EndNote, RefWorks, Bibtex, RefMan and WenXianWang). Once you select one of these, an additional link appears in every search result which allows you to export the citation (usually with a single click).

TH: What are Library Links?

AA: For libraries that make their resources available via a link resolver, we can include a link for their patrons to these resources as a part of the Google Scholar search results. On-campus users at participating schools will see these additional links in Google Scholar search results which facilitate access to their library's resources. These links lead to the library's servers which, in turn, direct them to the full-text of the article.

TH: How is the Library Links program doing?

AA: The Library Links program is doing very well. We have over a thousand institutions and consortia participating; with consortial participation, it is hard to determine the exact number of libraries. Library Links see significant use with roughly 10-20% click through, which is quite high for any given link in a large set of search results. We believe this program is quite successful in helping researchers discover and utilize the wealth of resources that libraries have licensed for them.

TH: I have heard that you are working with Ingenta and EBSCO on some form of linking from Google Scholar results. Can you explain what you are doing and why?

AA: The Library Links program requires participating libraries to have a link resolver. However, only a small fraction of libraries worldwide have link resolvers, and this number is growing fairly slowly even though Ex Libris provides ScholarSFX, a limited-feature link resolver as a free and hosted solution. This leaves a large number of researchers unable to take advantage of what their libraries have licensed as a part of their normal search workflow. We're trying to figure out ways to solve this problem. As a part of this, we're working with providers including Ingenta and EBSCO. This effort focuses on libraries that don't have link resolvers. Participating libraries that do have link resolvers should see no change.

TH: Have you noticed any problems with Library Links or how they are being used?

AA: We've noticed that some libraries don't specify all their IPs as a part of their Library Links information. In some cases, a significant fraction of the IPs belonging to the institution are not being included. This isn't great as users from these IPs don't see Library Links in their search results and therefore aren't able to discover and take advantage of their library resources. As I mentioned, when these links appear, they are used quite a bit. I strongly encourage participating libraries to check the set of IPs they are currently specifying, and update them if necessary.

TH: Can you tell us something about the Library Search program? How does it differ from Library Links?

AA: The Library Links program allows patrons of a given library (or a small number of known libraries) to take advantage of available resources. The Library Search program, on the other hand, allows users to discover which library or libraries have the literature they are looking for. We work with union catalogs worldwide to implement this feature, which appears as an additional link within each search result. We automatically select a union catalog based on the user's location (e.g., users from Sweden should see links to LIBRIS while users from the U.S. should see links to Open WorldCat).

TH: I have heard that you are also integrating union catalog results in Google Book Search, but I don't see them when I query Google Book Search. Can you tell us more? How does the integration work?

AA: We currently include union catalog results for selected Google Book Search queries. These queries are algorithmically selected based on several factors, including the degree of match for an entry, fields of match

(e.g. author, title), distribution of matches for all results for the query, and so on. The challenging issue here is the huge disparity in size between digitized books and metadata records, as well as the difference in user experience between them. The selection algorithm is evolving as we learn more about how this feature is used. You can try this feature out now by going to the [Advanced Book Search](#) page in Google Book Search and selecting the 'Library catalogs' radio button before performing your search.

TH: How well is the Library Search program working? How many union catalogs are currently participating?

AA: The Library Search program is doing quite well. Currently, about 20 union catalogs are participating. Together, they include libraries in over 35 countries. While not as convenient as links that provide online access, these links are still quite popular among users since they are often the only mechanism users have for locating a potentially accessible copy of a book.

TH: Why don't you provide a list of journals and/or publishers included in Google Scholar? Without such information, it's hard for librarians to provide guidance to users about how or when to use Google Scholar.

AA: Since we automatically extract citations from articles, we cover a wide range of journals and publishers, including even articles that are not yet online. While this approach allows us to include popular articles from all sources, it makes it difficult to create a succinct description of coverage. For example, while we include Einstein's articles from 1905 (the "miracle year" in which he published seminal articles on special relativity, matter and energy equivalence, Brownian motion and the photoelectric effect), we don't yet include all articles published in that year.

That said, I'm not quite sure that a coverage description, if available, would help provide guidance about how or when to use Google Scholar. In general, this is hard to do when considering large search indices with broad coverage. For example, the notes and comparisons I have seen about other large scholarly search indices (for which detailed coverage information is already available) provide little guidance about when to use each of them, and instead recommend searching all of them.

TH: Some librarians consider Google Scholar's interface too limited for sophisticated researchers. Do you plan to provide more options for manipulating or narrowing search results?

AA: Our experience as well as user feedback indicates that Google Scholar is widely used by researchers of all levels of sophistication -- from laypersons to leading experts. This is not surprising. LibQual's study of use of search habits of undergrads, graduate students and faculty members ([presentation available here](#)) shows that all three groups prefer general search engines with broad coverage and do so roughly with the same frequency.

Regarding options for narrowing and manipulating results, we do provide some on the [advanced search page](#). However, we have found that other than time-based restrictions (to search papers from the last few years), none of these options see much use. More generally, we refine the user interface for Google Scholar based on how people actually use it. Instead of considering a laundry-list of features we may add, we consider a list of frequently-performed operations and see how well we support them. A long list of unrelated features wouldn't be of much use. This is not surprising. For example, few of the tools in a full-featured Swiss Army knife see much use over its entire lifetime.

TH: Some librarians would like to have an API interface to Google Scholar results so that they can include it in a metasearch. As you know, libraries cover much more than scholarly literature and it is important to have one interface to search all resources that patrons have access to.

AA: This is indeed a request we've heard several times. The key issue with this, however, is that doing an effective metasearch is extremely hard. It's quite difficult to merge ranked results from different search interfaces with different ranking algorithms. This is difficult even when all the search indices being metasearched belong to the same organization (e.g., the different search services Google provides); for metasearch over diverse collections with different formats and characteristics, the situation is even more complicated. Nevertheless, I agree that being able to search over all material from a single search box is important. We will be happy to explore possibilities with interested folks from the library world. Feel free to contact us at scholar-library@google.com.

TH: I have heard that some institutions would like to have a version of Google Scholar for their patrons that is

limited to material available in their library or licensed for online use. They feel that finding articles they can't get frustrates users. What are your thoughts on this?

AA: I believe discovery needs to be universal -- researchers need to be able to find all that has been done in their area of interest. Believe me, I learned this the hard way. Much of scholarly endeavor is learning what has been done and building on it. You can't build on something you don't know about. Regarding the frustration, this is indeed an issue, but consider the alternative; it's much better to be frustrated than to be in the dark about what's out there.

TH: Any final words for our readers?

AA: Libraries are fabulous repositories of intellectual wealth. And they often are a major mechanism for leveling the playing field for those with limited resources. As someone who grew up on a small university campus in a developing country, I have a keen appreciation of what libraries can do to open new worlds for their patrons. I believe we (libraries and search engines) have an unprecedented opportunity to help users discover and take advantage of this wealth. Our Library Links and Library Search programs represent a good start. There is much more that we can do together. I invite all our readers to join us in this endeavor.