Google Scholar Revisited


Google Scholar had its debut in November, 2004. Although it is still in beta version, it is worth to revisit its pros and cons as changes took place in the past three years both in the content and the software of Google Scholar – for better or worse. Its content grew significantly - courtesy of more and more academic publishers and database hosts opening their digital vaults to allow the crawlers of Google Scholar to collect data from and index the full text of millions of articles from academic journal collections, and scholarly repositories of preprints and reprints. The Google Books project also gave a massive and valuable boost to the already rich and diverse content of Google Scholar. The dark side of the growth is that significant gaps remained for top ranking journals and serials, and the number of duplicate, triplicate and quadruplicate records for the same source documents (which Google Scholar cannot detect reliably) increased.

While the regular Google service does an impressive job with mostly unstructured Web pages, the software of Google Scholar keeps doing a very poor job with the highly structured and tagged scholarly documents. It still has serious deficiencies with basic search operations, does not have any sort options (beyond the questionable relevance ranking). It recklessly offers filtering features by data elements which are present only in a very small fraction of the records (such as broad subject categories), and/or are often absent and incorrect in Google Scholar even if they are present correctly in the source items. These include non-existent author names, which turn out to be section names, subtitles, or any part of the text, including menu option text which has nothing to do with the document or its author. This makes F. Password not only the most productive, but also a very highly cited author. Page numbers, the first or second segment of an ISSN, or any other four digit numbers are often interpreted by Google Scholar as publication years due to its artificial unintelligence. As a consequence, Google Scholar has a disappointing performance in matching citing and cited items. Its hit counts and citation counts remain highly inflated, defying the most basic plausibility concepts when reporting about documents from the 1990s citing papers to be published in 2008, 2009 or even later in the 21st century.

In spite of the appalling deficiencies and the shoddiness of its software, the free Google Scholar service is of great help in the resource discovery process and can often lead users to the primary documents in their library in print or digital format and/or to open access versions of papers which otherwise would cost more than $30-$40 a piece through document delivery services. Google Scholar can act at the minimum as a free, huge and diverse multidisciplinary I/A database or a federated search engine with limited software capabilities, but with the superb bonus of searching incredibly fast the full text of several million source documents. However, using it for bibliometric and scientometric evaluation, comparison and ranking purposes can produce very unscholarly measures and indicators of scholarly productivity and impact.
Background and literature

On the 3rd anniversary of Google Scholar, I give a summary of the pros and cons of Google Scholar, focusing on the increasingly valuable content and on the decreasingly satisfactory software features which must befuddle searchers, and ought to be addressed by the developers. I discuss here Google Scholar from the perspective of some of the traditional database evaluation criteria that have been used for decades (Jacso, 1998). I complement this paper with an unusually long bibliography of some of the most relevant English language articles by competent information professionals. For many of them I provided the URL of an open access prep-print or reprint version, or original version published in an open access journal, to offer readers convenient access to look into the papers and get the opinion of the authors. Having re-read these papers in preparing for this review was a great pleasure even when my opinion did not agree with that of the reviewers. The balance of pro and con arguments and evidentiary materials presented by competent information professionals has been rewarding and motivated my creating this bibliography. It does not include references to papers which are dedicated to the citation counts of articles as presented by Google Scholar. These will be provided in follow-up papers which discuss the strengths and weaknesses of using Scopus, Web of Science and Google Scholar to determine the Hirsch-index and derivative indexes for measuring and comparing research output in a quantitative way.

After the launch of Google Scholar it got much attention, just as anything gets a lot of ink that relates to Google, Inc. Within the first few months of its debut, there were a number of reviews in open access Web columns (Price, 2004; Jacso, 2004; Goodman, 2004; Gardner & Eng, 2005, Abram, 2005; Tenopir, 2005), and three Web blogs were launched dedicated to Google Scholar (Sondemann, 2005, Giustini, 2005), or partially dedicated (Iseid, 2006).

These were followed by reviews in traditional publications (Jacso, 2005a, Myhill, 2005, Notess, 2005, O’Leary, 2005, Giustini and Barsky, 2005; Noruzi, 2005; Adlington and Benda, 2006; Cathcart and Roberts, 2006) focusing on the content and software aspects of Google Scholar. These were well complemented by a number of essays, editorials and surveys, pondering about the acceptance, use, promotion and “domestication” of Google Scholar as one of the endorsed research tools for students and faculty in academic institutions (Kesselman and Watsen, 2005; Price, 2005; Anderson, 2006; Gorman, 2006; Mullen and Hartman, 2006; Friend, 2006, Hamaker and Spry, 2006, York, 2006; Helms-Park et al., 2007; Schmidt, 2007, Taylor, 2007).

As Google Scholar became more intensively used, several research papers started to put it into context by comparing Google Scholar’s performance with a single database (Schultz, 2007), federated search engines (Felter, 2005, Giustini and Barsky, 2005; Chen, 2006; Sadeh, 2006; Donlan and Cooke, 2006; Haya et al., 2007; Herrera, 2007), citation-enhanced databases such as Web of Science and/or Scopus (Bauer and Bakkalbasi, 2005; Jacso, 2005b; Jacso, 2005c; Yang and Meho, 2006; Norris and Oppenheim, 2007), or with a mix of these and traditional scholarly indexing/abstracting databases (White, 2007).

There is increasing specialization in researching Google Scholar, applying the traditional database evaluation criteria such as size, timeliness, source type and
especially breadth of journal coverage, (Jacso, 1997) in a consistent manner in the context of a very non-traditional database which piggybacks on other sources rather than creating its own (Wleklinksi, 2005, Vine, 2005, Vine, 2006, Neuhaus et al., 2006; Pomerantz, 2006; White, 2006; Mayr and Walter, 2007; Walters, 2007). The recent incorporation of books in Google Scholar from Google Book Search (which after a poor debut with deficient software features, turned around and introduced within a month a far more sophisticated software than Google Scholar in three years), spawned useful research (Hauer, 2006; Lackie, 2006; Goldeman and Connolly, 2007). So did the only good new software feature of Google Scholar to lead the users to the full text digital source document in the users’ library through Open-URL resolvers (Grogg and Ferguson, 2005; O’Hara, 2007; Legace and Chisman, 2007).

There is one additional research area where Google Scholar will play an important role, its use for bibliometric and scientometric evaluation of the performance of researchers, which is such a complex issue that deserves to be discussed in a separate paper, with its own rich set of references.

THE PROS
Most of the pros relate to the content part of Google Scholar, from different angles, including coverage, variety in source and journal base, size, currency.

Journals coverage
The source base of Google Scholar has been considerably enhanced since its debut, as every scholarly publishers want to be a part in the Google Universe. The source base increased also in quality by full-text indexing of thousands of additional academic journals of great importance from the sites of the publishers, rather than just indexing bibliographic data and abstract from I/A databases. The two most important journal publishers that started to cooperate with Google Scholar are Elsevier and the American Chemical Society. Although only a tiny part were indexed so far by Google Scholar from these publishers’ important digital collections of more than 7 million (Elsevier) and 0.75 million (ACS) items, their shares are expected to increase at a rapid pace, once the Google Scholar spiders’ are sent to their routes. It was an excellent idea to add book records to Google Scholar, primarily from the Google Books project. It is a huge advantage as books are barely present even as an indexing/abstracting record, let alone as a completely indexed full text item (for searching, not viewing) in most of the other multidisciplinary mega-databases (except for the also free and outstanding Amazon.com site). In preparing for a tutorial session in Vietnam, it was impressive to find 27 books in/through Google Scholar each of which had numerous passages about or references to the so-called scholar gentry class. This is the type of casual digital book use that the late Frederick Kilgour, the founder of OCLC envisioned – more than 20 years ago, when he was already in his early 70s.

Geographic and language coverage
The geographic and language coverage of Google Scholar is also impressive and genuine. It is a typical limitation of even the subscription-based scholarly databases that they often almost exclusively cover only English language sources, predominantly published in the U.S., U.K., Australia and Canada (in which case French language documents are also covered). I don’t blame the commercial database publishers for this as they were not created on the same principles as the U.N. or UNESCO. They have to spend their money on processing documents which are of interest and understandable for the majority of scholars, their primary customers.

The Google Scholar service does not have the ever-increasing costs of subscription and human processing of the scholarly print publications. It has free access to practically any scholarly digital document collection it wants, and wisely decided to index (by software) important Spanish, Portuguese, German, Japanese, Chinese, Korean and Russian language collections of academic works. While the latter four are of no help for me, the former three are, indeed, and are worth the extra mental effort to read in the native language, as there are several sources in my areas of specialization where researchers in Germany, Austria, in the Iberian peninsula, in Central and South America, especially in Spain and Brazil, that publish only in German, Spanish and Portuguese.

I have avoided to refer to the actual size of Google Scholar and its subsets as it is impossible to come up with a realistic number, or even with an estimate for the number of records in the database, or in the Canadian subset, or in the language subsets.

Digital repositories
The coverage of digital repositories—even if far from complete—is already a great asset, especially for physics, astrophysics, medicine, economics, and computer and information sciences and technology. But the use of such full text repositories still could be significantly improved. For example, only about a quarter of the open access PubMed Central (PMC) items are directly available in Google Scholar. True, there are records in Google Scholar—from other sources, such as cababstractsplus.org—for many more of the 620,000 full text documents reposited in PMC.

It would be, however, essential to index the source documents, and give them priority in displaying the result list clearly marking them as open access, instead of giving undeserved prominence to the British Library document delivery service (BL Direct) which is more than happy to charge an arm and a leg for document delivery even when the open access paper is just a click away from the user. Just as quickly as Google Scholar can figure out if a journal is available for article delivery through the British Library, it could figure out if is available free of charge from runs of open access issues of the journal. The same is true for the open access full text subset of the National Transportation Library (which has, for example, more than 100 documents about transportation-related terrorism). In sharp contrast, Google Scholar has only a dozen source documents indexed and made available from that site in their full glory.
While praising the broad content coverage of Google Scholar, it must be noted that there are still huge gaps in the full text indexing of the most important serial publications that was mentioned in the original review (Jacso, 2005). For example, less than 17% of the 430,500 documents at the nature.com web site were indexed by Google Scholar directly from that site (which includes not only Nature magazine but many other journals of the Nature Publishing Group). True, many more than 17% of them have a record in Google Scholar but many of these are just citation records with minimal information.

Indexing/abstracting records
It is good that there millions of records from good indexing/abstracting databases, for documents for which digital full text is not (yet) available. However, Google Scholar should have used the unique privilege granted by thousands of scholarly publishers of getting unfettered access to, crawl and index the full text of the primary documents, rather than just the ersatz records, often redundantly through several indexing/abstracting databases.

Size
I usually start the content review by determining the size of the database, and its distinct subsets. It is essential for researchers to know how many records there are in Google Scholar in total, and/or in, say, English or Spanish, which journals are covered from what publishers for what time span, but its developers “take the Fifth” when asked about it or about any factual features of the database (such as the number of journals, publishers, foreign language materials, articles, conference papers, reports, books covered at least in general). My various “sizing up” queries don’t work as the results are so absurd and/or capricious that it would be irresponsible to report them.

The only good new features in the software are the Library Links and Library Search options. These will inform the users if their library offers access to the document in question. If your library signed up (and provided data about its digital journal holdings) to Google Scholar this would work automatically (if Google Scholar is invoked from the library or a computer with authenticated IP address, or remotely through the library, after the appropriate log-in process). The Library Search option for books works if the library is an OCLC member. It is to be noted that the [BOOK] label in the Google Scholar result lists often refer to a review of or blurb about the book, not to the book itself.

THE CONS
Practically, all the major negative traits of Google Scholar are caused by or relate to software issues. As I indicated above, it is impossible even to guess the size of the database because of the elementary problems with the software.

Innumeracy
It speaks volumes about the limitations of the software that when using the query term <the> (the most commonly occurring English word), Google Scholar yields a hit count of over 1.5 billion records, whether you are using it with or without the + sign.
or surround it by double quotation marks (as it is supposed to be a stop word without these signs, but apparently it isn’t). I don’t believe this hit count to be true, but that is not the point here.

If you add (out of curiosity) the letter *a* in an OR relationship, the result set should increase by picking up records for foreign language source documents which use the letter *a* as the definite article and/or a preposition. In the extreme case if all the English language records had the latter *a* as the indefinite article or part of terms such a blood type A, personality A, grade A, etc., the number of hits would not increase but would remain the same.

But in Google Scholar the OR operator decreases the result set to less than 1% of the original set, and makes George Boole rolling in his grave. The regular Google engine does not do this nonsense. Some may feel lucky (rather than befuddled) that although both search terms were purportedly excluded from the search (as the message shows), Google Scholar still could come up with nearly 14 million hits – without using the + sign or the double quotation mark. Actually, it shows only 1,000 hits at most, for any query, so it can claim any number above 1,000 without the burden of proof. If gamblers could bluff in casinos without the burden of showing their cards for the blackjack dealer, there would be many more instant millionaires than at the GooglePlex.

This has been a problem from the beginning. The enhancement of the content has not been matched by improvements in the software. The software does not reflect at all, for example, the specialties of the fully indexed books. The template in the advanced mode still refers to articles written by, articles published in, articles published between, and articles in subject areas.
As for subject areas, they should not be really used as filters. When entering the search for any documents, with the word Vietnam in the title, and the radio button for all subject areas is left turned on, Google Scholar reports 135,000 hits, an impressively high number. When sending the query through the advanced template Google Scholar inserts two spaces in front of the search term. If you change it to one, the result will go up to 137,000, if you eliminate both spaces the result set will revert to 135,000 items. This is not true for the field-specific searches, such as author, title, journal name. This will be the least enigmatic part of the search process thanks to the weird logic of Google Scholar.

Selecting one check-box at a time for filtering by the first subject group, then the second, the third, etc. will produce cumulative subsets. After the last subject group the aggregate of the seven subject categories will produce a set of 20,500 records. This is less than 15% of the original set, meaning that 85% of the items for this topic are not assigned to any of the subject groups.
Figure 4. Selecting each listed categories the set decreases by 85%

Much more surprisingly, when the query is expanded by adding the word Vietnamese to the query without any filtering, the result will shrink to 46,100 items, 34% of the single word query.
Figure 5. Expanding the query will drastically shrink the result set

More oddly, restricting the search to the seven listed subject groups will increase the result set to 105,000. Activating the Search in all subject areas radio button will report a set size of 43,200 (not shown here because any logic breaks down here, and only the first 1,000 items will be listed by Google Scholar anyhow).

Figure 6. Restricting the query to predefined subject categories will more than double the set

The publication year limiters behave in an equally odd way. It is madness and there is no method to it. Limiting the initial set with Vietnam in the title to the publication year range 1435-2008, to accommodate the first possible English language transliteration of the Vietnamese word for the name of the country to publications which will be published the next year (I write this in mid-November, 2007) yields 20,200 hits. Limiting the search to 1960-2008, i.e. to a more than 500 years shorter time span, increases the set to 20,600 items. The fact that many records in any sample would not have the publication year data element, or Google Scholar would not recognize it even when it is right under its nose, does not “justify” this weird logic. There is not a word about this serious limitation in the cheery help file.
Illiteracy
These were problems of innumeracy, but there are many problems that can be classified as problems of illiteracy in the software. When the two come together in certain searches the chaos gets serious. Google Scholar has lethal deficiencies in distinguishing author names from other parts of the text using its parsing algorithm. Apparently, the developers are not aware of it, or don’t care about it.

After seeing left and right author names like F. Password, V. Findings, N. Vietnam, S Vietnam, it was surprising to notice one of the new software features of Google Scholar, the cluster of authors related to the user’s query as explained in the help file. My test search shows the suggested authors from a set of purportedly 2,9110,000 records on the topic of risk factor evaluation with the following names: P Population, R Evaluation, M Data, R Findings and M Results. Google Scholar flaunts its software deficiencies, and does not provide any hints about the limitation of the software.

Figure 7. The shorter the time span the higher the hit count
How do I search by category?

From the Advanced Search page, you can search for scholarly literature within seven broad areas of research. Simply check the boxes for the subject areas you're interested in searching.

Why are there author names on the left hand side of my results page?

We automatically suggest authors related to your query – just click on an author's name and you'll see their papers. Finding authors who publish on the topics you're interested in is often a great way to get better acquainted with a field and discover related work you may not have found otherwise.

Figure 8. Odd list of recommended authors in the side bar, and a cheery help file

The extent of wrong author names is well above hundreds of thousands, and often they deprive the real authors from getting credit for some of their paper (including highly cited papers) - and thus from getting a decent h-index.
The upcoming issues will look at the theory and the practice of determining the h-index in general, and in Google Scholar, Scopus and Web of Science in particular.

References


