Is Google Scholar useful for Bibliometrics? A Webometric Analysis

Isidro F. Aguillo¹

¹ isidro.aguillo@cchs.csic.es The Cybermetrics Lab. CSIC, Albasanz, 26-28, Madrid 28037 (Spain)

Abstract

Google Scholar, the academic bibliographic database provided free-of-charge by the search engine giant Google, has been suggested as an alternative or complementary resource to the commercial citation databases like Web of Knowledge (ISI/Thomson) or Scopus (Elsevier). In order to check the usefulness of this database for bibliometric analysis, and especially research evaluation, a novel approach is introduced. Instead of names of authors or institutions, a webometric analysis of academic web domains is performed.

The bibliographic records for 225 top level web domains (TLD) and 19,240 university institutional web domains has been collected from the Google Scholar database. About 63.8% of the records are hosted in generic domains like com or org, confirming that most of the Scholar data come from large commercial or non-profit sources. One third of the other items (10.6% from the global) are hosted by the 10,442 universities that have at least one record in Scholar. The individual analysis show that universities from China, Brazil, Spain, Taiwan or Indonesia are far better ranked than expected. In some cases, large international or national databases, or repositories are responsible for the high numbers found. However, in many others, the local contents, including papers in low impact journals, popular scientific literature, and unpublished reports or teaching supporting materials are clearly overrepresented.

Google Scholar lacks the quality control needed for its use as a bibliometric tool; the larger coverage it provides consists in some cases of items not comparable with those provided by other similar databases.

Introduction

Google Scholar (http://scholar.google.com/) is the database developed by Google Inc. to provide access to the world scholarly literature. The platform use both the academic records from its main search engine but also many other sources including commercial, non-profit, institutional or individual bibliographic databases. Google Scholar was introduced in November 2004 and it is still in beta version, although several major changes have occurred since then, as the coverage has been increased considerably, new formats are available (including patents and legal opinions, theses, books, abstracts and articles) and additional operators are being added (Mayr & Walter, 2007; Jacsó, 2008; Torres-Salinas, Ruiz-Pérez & Delgado López-Cozar, 2008).

There are several relevant features that explain the success of Scholar. It is provided free-ofcharge, offering perhaps one of the largest scientific bibliographic databases. It is build from combining an undisclosed number of very large databases, whose contents are not available to the public web, plus those belonging to the so-called invisible web and the academic related web documents from the huge Google search engine and it includes citations to the items. So Google Scholar is comparable to the other two large multidisciplinary citation databases, Web of Knowledge (edited by ISI/Thomson) and Scopus (developed by Elsevier), both of which are commercial and hugely priced sources and key tools for the analysis and evaluation of scientific activity and results.

The growing interests in this database by the scientometric community was fuelled in part by the launch of *Publish or Perish* (http://www.harzing.com/pop.htm), a free program to automatically recover records from the Scholar web gateway that also provides a series of basic and sophisticated bibliometric indicators like the h-index family (Harzing & van der Wal, 2008a,b).

Preliminary global analyses of Google database uncovered several problems and shortcomings (Bar-Ilan, 2009; Jacsó, 2008; Kousha & Thelwall, 2008; White 2006). As the records came from very different sources, formal integration was not possible in many cases,

a situation much more worsened because of the lack of control of its contents. The result is a very noisy database that requires a lot of difficult and time consuming cleaning effort to obtain useable information, especially for evaluation purposes. Several authors (Bar-Ilan, 2010; García-Pérez, 2010; Jacsó, 2010; Li *et al*, 2010; Meho & Yang, 2007; Mikki, 2010) published comparative analyses with the other two citations databases (WoK and Scopus), but also other specialised databases, stating the overlapping of contents and the large coverage of Scholar, both in the number of papers provided but also in the typology, as conference presentations, internal reports, unpublished drafts or teaching-supporting material was available.

Apart from methodological issues, it looks like a consensus has been agreed to consider Scholar as a (cheap) complement tool to the other citation databases in the bibliometric studies.

However, citation databases are relevant not only because they offer precisely *citations*, but because there is a very strict quality of the journals indexed (those in the elite of the Bradford nucleus, blurred sometimes due to commercial policies). It is also taken for granted that Elsevier (Scopus) and ISI/Thomson (WoK) use the same criteria, which is not clear. ISI's criteria are known, but not SCOPUS ones. For instance all of the Elsevier's journals are included systematically, which is not exactly a guarantee of quality.

Our hypothesis is that Google Scholar already covers items from low impact sources, a fact probably unnoticed because most analysis focuses on overlapped contents with the excellence-driven WoK and Scopus. If this is true, more caution should be taken when Google Scholar is used for evaluation purposes.

In order to check this hypothesis in a global scenario, a risky approach involving webometric methods is proposed. Instead of traditional institutional affiliations, the institutional web domains will be used for the analysis (Aguillo, 2009).

Methodology

The collection of data took place during August 2010, using the canonical address of Google Scholar (http://scholar.google.com) and the following syntax:

site:tld (example: *site:edu*) or *site:institutional domain* (example *site:cam.ac.uk*).

The filtering criteria applied were: "articles, excluding patents" and "at least summaries", in order to obtain the total number of items by national or institutional web domain. Contrary to the irregularity of the numbers obtained from Google, the academic database is stable and changes occur from updating with new records (size tends to increase after about two weeks).

For general analysis we identified 225 top level domains (TLD), including generic or international ones like com, org or net (gTLD) and national (country) ones, like es, fr or it (cTLD). USA is represented by the combined domains us (mainly local and state related), gov (federal gov), edu (mostly US universities) and mil (military). The total number of records obtained was over 86 million, which it is an overestimation as several copies of the same document can be hosted in different servers.

The second population analyzed is a list of university domains obtained from the July 2010 edition of the Catalogue of World Universities of the Ranking Web of Universities (http://www.webometrics.info/university_by_country_select.asp), probably the most complete and updated list of higher education institutions (HEIs) currently available. About 19,240 web domains were analysed, resulting in 10,442 HEIs with at least one item in Google Scholar. The total number of records is about 9 million for the whole university sector.

Results

The distribution by TLD domain is showed in Table 1. The combined gTLD, mostly consisting of profit (com) and not-for-profit (org) organizations amounts for the 64% of the

total number of records. Most of these organizations are probably from North America and Europe, being the well-known major publishers well represented, although the list of contributors has not been made public by Google. The geographic coverage of this section is probably reflecting the same biases as other major databases with similar sources.

The distribution by country domain also includes companies and other non-HEIs organizations, using the national suffix. Some of them are truly international (the larger European ones), but there are also others mainly operating from one country with international (com, org, net) domains. All the cTLDs combined represents about 36%, half of them in academic and research organizations (10.6 % in universities, Table 2).

Country	Domain	Items	%
gTLD	com,org,net,	54,862,451	63.79%
USA	edu,gov,us,mil	7,873,000	9.15%
China	cn	7,520,000	8.74%
France	fr	2,820,000	3.28%
Japan	jp	1,720,000	2.00%
Brazil	br	1,440,000	1.67%
Russia	ru	995,000	1.16%
Spain	es	907,000	1.05%
Taiwan	tw	752,000	0.87%
Germany	de	684,000	0.80%
Canada	ca	552,000	0.64%
South Korea	kr	481,000	0.56%
United Kingdom	uk	430,000	0.50%
Australia	au	399,000	0.46%
Italy	it	308,000	0.36%
Switzerland	ch	227,000	0.26%
Poland	pl	220,000	0.26%
Netherlands	nl	219,000	0.25%
Ukraine	ua	210,000	0.24%
Mexico	mx	203,000	0.24%
Costa Rica	cr	177,000	0.21%
TOTAL	225	86,010,880	

Table 1. Number	of items recovered from	Google Scholar for	the Top Level Domain	ıs, including
both	generic or international	(gTLD) and nation	al ones (August, 2010)	

Only the 54.3% of the universities have at least one record in Google Scholar. The average number of items of those HEIs represented in the database is 870, although with large differences between regions and countries. North America (USA & Canada) and Oceania (Australia & New Zealand) are hosting more than 1600 items as a mean (Table2).

Region	Universities	Items	Average	%
North America	2,407	3,936,623	1,635	43.34%
Europe	3,223	2,657,514	825	29.26%
Asia	2,615	1,225,026	468	13.49%
Latin America	1,809	1,033,097	571	11.37%
Oceania	92	152,317	1,656	1.68%
Arab World	194	40,677	210	0.45%
Africa	102	37,333	366	0.41%
TOTAL	10,442	9,082,587	870	

 Table 2. Distribution by regions of the number of universities, the items in Google Scholar for each university webdomain and the average size of records by institution (August, 2010).

However the most interesting data is the distribution by country. According to the total number of items in the universities of the country (Table 3), after the USA, Spain is ranked the second, Brazil the third and Taiwan the fourth. Costa Rica and Indonesia are positioned among the top ten. There is a great diversity regarding the relative contribution of the academic sector that can be explained by national differences: existence of large research councils (France) or academy of sciences (China, Russia), if there are strong open access mandates and policies (Spain), and if they host "international" huge repositories (Mexico, Costa Rica).

Country	Total	University	%
USA (edu, gov, us,mil)	7,873,000	3,711,305	47.10%
Spain	907,000	717,078	79.10%
Brazil	1,440,000	492,525	34.20%
Taiwan	752,000	329,936	43.90%
Japan	1,720,000	319,757	18.60%
Germany	684,000	282,887	41.40%
Canada	552,000	225,277	40.80%
United Kingdom	430,000	207,082	48.20%
Costa Rica	177,000	167,444	94.60%
Mexico	203,000	165,604	81.60%
Indonesia	160,000	155,249	97.00%
Netherlands	219,000	148,376	67.80%
Russia	995,000	141,085	14.20%
China	7,520,000	133,241	1.80%
Australia	399,000	126,063	31.60%
Italy	308,000	125,828	40.90%
France	2,820,000	104,535	3.70%
Sweden	167,000	93,503	56.00%
Belgium	129,000	62,176	48.20%

 Table 3. Comparison between the total number of Scholar items and the combined number of items in all the universities sharing the same top level domain (August, 2010)

Aguillo

Switzerland	227,000	62,130	27.40%
South Korea	481,000	53,048	11.00%
Poland	220,000	51,820	23.60%
Ukraine	210,000	38,792	18.50%
Argentina	125,000	37,748	30.20%
TOTAL	86,010,880	9,082,587	10.60%

Further analysis requires checking the individual institutions as it is possible to identify the reasons for that prominence in several cases. For example, in Table 4, Harvard is hosting a large astronomical database, PSU is serving CiteSeerX, Rioja maintains the large Dialnet repository, Johns Hopkins is developing MUSE project, CATIE is the central organization for a large agricultural database and there are other databases/repositories linked to many of the rest of universities (Bibliodoc-Complutense; Redalyc-UAEM or Brazilian thesis-USP).

But this is not an easy task, and there is no explanation for many other institutions, especially those that have increased significantly the number of records in the period 2005-2009 (mostly Brazilian and Taiwanese universities). In the case of Taiwan, the number of records in Chinese suggests that many (but not all) of the contributions are of local origin.

University	Total	2000-04	2005-09
Harvard University	1,170,000	197,000	235,200
Pennsylvania State University	1,060,000	212,000	240,300
Universidad de La Rioja	422,000	99,100	140,000
Johns Hopkins University	214,000	41,700	100,700
Centro Agronómico Tropical de Investigación y Enseñanza	158,000	11,700	15,880
Universidad Complutense de Madrid	110,000	25,600	33,480
Universidad Autónoma del Estado de México	97,700	18,400	25,940
Universidade de São Paulo	71,500	3,620	10,140
Massachusetts Institute of Technology	67,000	10,200	17,060
National Taiwan University	64,400	17,800	27,370
University of Zagreb	63,200	7,870	19,330
Universidade Federal de Santa Catarina	44,100	3,590	6,168
Stanford University	39,800	4,920	6,180
Kyoto University	38,700	6,040	5,807
National Chung Hsing University	38,300	5,770	31,430
University of Michigan	33,600	3,160	5,549
Universidade Estadual de Campinas	33,200	3,410	5,830
University of Minnesota	33,200	5,990	9,130
Universidade Federal do Rio Grande do Sul	31,500	4,540	15,120
Masaryk University	30,800	2,340	6,007
Universidad Nacional Autónoma de México	30,100	2,110	9,425
National Central University	29,600	9,870	15,470
National Tsing Hua University Taiwan	28,300	7,000	7,440
University of British Columbia	28,000	2,240	3,322
University of Oslo	26,500	1,230	2,311
University of Nebraska Lincoln	26,500	4,840	5,180
National Cheng Kung University	25,500	6,800	15,170

University of Groningen	25,500	2,150	5,022
University of Tokyo	24,800	3,540	5,158
Utrecht University	24,400	2,290	3,734

Discussion and conclusions

The webometric approach is simpler and faster, but it is also significantly noisier, as the university webdomains could host papers not authored by the institution's scholars. For example, the conference proceedings of events taking place in the institution are usually published in its webpages, when in fact a lot of those contributions were added by external scientists. Another source of foreign texts came from teaching supporting material (seminal papers, book chapters) offered by local scholars (sometimes even without legal permission of its original authors). In many cases these are highly cited papers that are relevant contributions or comprehensive reviews. Finally, a few of the most important field repositories, with hundreds of thousands papers coming from all over the world, are using servers with domain names of the universities (CiteSeerX in Pennsylvania State University: http://citeseerx.psu.edu/). This fact is evident in many of the top ranked institutions showed in Table 4.

A complete different source of problems is the web publication of informal material, drafts papers, unpublished reports or academic handouts being incorporated to Scholar database. Checking the contents of a few institutions appearing in Table 4, it is easy to obtain evidence of local magazines indexed cover-to-cover, institutional reports, chapters or books addressed to general audiences or XVII-XX centuries digitised works.

We do not attempt to segregate the full-text records from those providing only a summary. According to the business model of Google Scholar, most of the commercial database providers are only supplying the summaries, with a link to the publisher/distributor page where the full paper can be obtained at a price. According to the data in Table 3, the records hosted by all the universities only represent about 10% of the total, but probably most of them are full text documents, while many of the records under gTLD domains are mainly summaries.

Google Scholar has increased the volume of contents it indexed directly from the web, (sub)domains example following especially from (for this model http://repository.university.edu/) in the academic Webspace devoted to host the papers, presentations and other documents of the universities. In many cases Google identify in some way, or they are informed about, the repository web address, and then, using crawlers, Google scan these addresses and the contents are added to the main Scholar database. This is an important source of local documents that are published in the web without making quality distinction and even in larger numbers than those of other universities that only make available a small fraction of their (high impact) papers because they are not enforcing open access policies.

Google Scholar was not designed as a direct competitor to the other citation databases, being this extra feature (citation counts and links) mainly oriented to improve the searching experience. It is really a huge database and Google is clearly intending to enlarge its coverage, not only by adding additional sources but by collecting every type of scientific material available from the public web. The university webdomains are relevant sources, but there are many cases where there is no quality control not by the scholars or by Google.

Our suggestion is that the use of Google Scholar for bibliometric or evaluation purposes should be done with great care, especially regarding the items not overlapping with those present in the Scopus or WoK citation databases.

References

- Aguillo, I. (2009). Measuring the institution's footprint in the web. Library Hi Tech, 27 (4), 540-556.
- Bar-Ilan, J. (2007). Which h-index? A comparison of WoS, Scopus and Google Scholar. *Scientometrics*, 74(2), 257–271.
- Bar-Ilan, J. (2009). A Closer Look at the Sources of Informetric Research. *Cybermetrics*, 13: Paper 4. http://www.cindoc.csic.es/cybermetrics/articles/v13i1p4.pdf
- Bar-Ilan, J. (2010) Citations to the "Introduction to informetrics" indexed by WOS, Scopus and Google Scholar. *Scientometrics*, 82 (3), 495-506
- Beel, J. & Gipp, B. (2010). Academic search engine spam and Google Scholar's resilience against it. *Journal of Electronic Publishing*, 13 (3)
- García-Pérez, M.A. (2010). Accuracy and completeness of publication and citation records in the Web of Science, PsycINFO, and Google scholar: A case study for the computation of h indices in psychology. *Journal of the American Society for Information Science and Technology*, 61 (10), 2070-2085
- Harzing, A. & van der Wal, R. (2008). Google Scholar as a new source for citation analysis. *Ethics in Science and Environmental Politics*, 8(1):61–73
- Harzing, A. & van der Wal, R. (2008). A Google Scholar h-index for journals: An alternative metric to measure journal impact in economics and business. *Journal of the American Society for Information Science*, 60(1), 41–46.
- Jacsó, P. (2008). Google Scholar revisited. Online Information Review, 32 (1), 102-114
- Jacsó, P. (2010). Savvy searching pragmatic issues in calculating and comparing the quantity and quality of research through rating and ranking of researchers based on peer reviews and bibliometric indicators from Web of Science, Scopus and Google Scholar. *Online Information Review*, 34 (6), 972-982
- Kousha, K. & Thelwall, M. (2008). Sources of Google Scholar citations outside the Science Citation Index: A comparison between four science disciplines. *Scientometrics*,74(2), 273–294
- Li, J., Burnham, J.F., Lemley, T. & Britton, R.M. (2010). Citation analysis: Comparison of Web of Science, Scopus, Scifinder, and Google Scholar. *Journal of Electronic Resources in Medical Libraries*, 7(3), 196-217
- Mayr, P. & Walter, A-K. (2007). An exploratory study of Google Scholar. *Online Information Review*, 31 (6), 814-830.
- Meho, L. & Yang, K. (2007). Impact of data sources on citation counts and rankings of LIS faculty: Web of Science vs. Scopus and Google Scholar. *Journal of the American Society for Information Science and Technology*, 58: 2105–25.
- Mikki, S. (2010). Comparing Google Scholar and ISI Web of Science for earth sciences. *Scientometrics*, 82 (2), 321-331
- Torres-Salinas, D., Ruiz-Pérez, R. & Delgado-López-Cózar, E. (2008). Google Scholar como herramienta para la evaluación científica. *El profesional de la información*, 18 (5), 501-510
- White, B. (2006). Examining the claims of Google Scholar as a serious Information Source. *New Zealand Library & Information Management Journal*, 50 (1): 11-24