Journal of Informetrics 12 (2018) 1160–1177

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Regular article

Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories

Alberto Martín-Martín[a,*], Enrique Orduna-Malea[b], Mike Thelwall[c], Emilio Delgado López-Cózar[a]

[a] Facultad de Comunicación y Documentación, Universidad de Granada, Granada, Spain
[b] Universitat Politècnica de València, Valencia, Spain
[c] Statistical Cybermetrics Research Group, School of Mathematics and Computer Science, University of Wolverhampton, Wolverhampton, UK

ARTICLE INFO

Article history:
Received 15 August 2018
Received in revised form 30 August 2018
Accepted 3 September 2018

Keywords:
Google Scholar
Web of Science
Scopus
Bibliographic databases
Academic search engines
Coverage
Citation analysis
Unique citations
Citation overlap
Bibliometrics
Scientometrics

ABSTRACT

Despite citation counts from Google Scholar (GS), Web of Science (WoS), and Scopus being widely consulted by researchers and sometimes used in research evaluations, there is no recent or systematic evidence about the differences between them. In response, this paper investigates 2,448,055 citations to 2299 English-language highly-cited documents from 252 GS subject categories published in 2006, comparing GS, the WoS Core Collection, and Scopus. GS consistently found the largest percentage of citations across all areas (93%–96%), far ahead of Scopus (35%–77%) and WoS (27%–73%). GS found nearly all the WoS (95%) and Scopus (92%) citations. Most citations found only by GS were from non-journal sources (48%–65%), including theses, books, conference papers, and unpublished materials. Many were non-English (19%–38%), and they tended to be much less cited than citing sources that were also in Scopus or WoS. Despite the many unique GS citing sources, Spearman correlations between citation counts in GS and WoS or Scopus are high (0.78-0.99). They are lower in the Humanities, and lower between GS and WoS than between GS and Scopus. The results suggest that in all areas GS citation data is essentially a superset of WoS and Scopus, with substantial extra coverage.

1. Introduction

The launch of Google Scholar (GS) in November of 2004 brought the simplicity of Google searches to the academic environment, and revolutionized the way researchers and the public searched, found, and accessed academic information. Until that point, the coverage of academic databases depended on lists of selected sources (usually scientific journals). In contrast, and using automated methods, Google Scholar crawled the web and indexed any document with a seemingly academic structure. This inclusive approach gave GS potentially more comprehensive coverage of the scientific and scholarly literature compared to the two major existing multidisciplinary databases with selective journal-based inclusion policies, the Web of Science (WoS) and Scopus (Orduna-Malea, Ayllón, Martín-Martín, & Delgado López-Cózar, 2015).

Although citation data in Google Scholar was originally intended to be a means of identifying the most relevant documents for a given query, it could also be used for formal or informal research evaluations. The availability of free citation data in

* Corresponding author.
E-mail address: albertomartin@ugr.es (A. Martín-Martín).

https://doi.org/10.1016/j.joi.2018.09.002
1751-1577/

Table 1
Results of studies that analysed unique and overlapping citations in GS, WoS, and Scopus.

Study	Sample	N citations	% only GS	% only WoS	% only Scopus	% only GS & WoS	% only GS & Scopus	% only WoS & Scopus	% GS & WoS & Scopus	% GS (all cit.)	% WoS (all cit.)	% Scopus (all cit.)	% WoS cit. in GS	% Scopus cit. in GS
Bakkalbasi, Bauer, Glover, & Wang (2006)	50 journal articles covered in JCR Oncology	614	13	7	12	4	5	28	31	53	70	76	215/431 = **50%**	220/469 = **47%**
	50 journal articles covered in JCR Physics, Cond. Matter	296	17	20	8	9	3	22	21	50	72	54	84/212 = **40%**	72/162 = **44%**
Yang & Meho (2007)	Scientific production of two Library & Information Science (LIST) researchers	385	10	23	6	10	7	18	25	52	77	57	137/295 = **46%**	124/218 = **57%**
Meho & Yang (2007)	1,457 articles published by 25 LIS researchers	5,285	48	Only (WoS or Scopus): 21		GS-(WoS or Scopus): 31		NA	NA	79	38	44	% (WoS or Scopus) cit. in GS 1,629/2,733 = **60%**	
Kousha & Thelwall (2008)	262 WoS-covered Biology journal articles	1,554	17	28		55				72	83		847/1288 = **66%**	
	276 WoS-covered Chemistry journal articles	729	8	62		30				38	92		218/668 = **33%**	
	262 WoS-covered Physics journal articles	1,734	36	24	NA	40	NA			76	64	NA	690/1111 = **62%**	NA
	82 WoS-covered Computing journal articles	3,369	67	14		19				86	33		632/1117 = **57%**	
	Total WoS-covered journal articles (882)	7,386	43	24		32				76	57		2387/4184 = **57%**	
Jacimovic, Petrovic, & Zivkovic (2010)	158 articles published in Serbian Dental Journal	249	58	4	6	1	2	15	15	76	34	39	39/85 = **46%**	43/94 = **46%**
Bar-Ilan (2010)	Book "Introduction to Informetrics" by L. Egghe and R. Rousseau	397	27	12	2	6	5	9	39	77	66	55	177/259 = **68%**	174/218 = **80%**
Lasda Bergman (2012)	5 top journals in the field of Social Work	4,308	44	5	8	2	8	12	22	76	41	50	1042/1741 = **60%**	1285/2126 = **60%**
de Winter, Zadpoor, & Dodou (2014)	Garfield, E. (1955). Citation indexes for science. *Science*, 122(3159), 108-111.	1,309	33	41	NA	35	NA			68	76	NA	453/606 = **75%**	NA
Rahimi & Chandrakumar (2014)	2,082 WoS-covered articles in General and Internal Medicine	62,900	29	10	11	2	9	8	31	71	51	59	20532/31778 = **65%**	25180/37272 = **68%**
Moed, Bar-Ilan, & Halevi (2016)	Articles published in 12 journals from 6 subject areas	6,941	47	NA	6	NA	47	NA	NA	94	NA	53	NA	3246/3651 = **89%**

NA = not analysed in the study.
Cells with more intense background color represent higher percentages of citations within the same sample of documents.

Google Scholar, together with the free software *Publish or Perish* (Harzing, 2007) to gather it made citation analysis possible without a citation database subscription (Harzing & van der Wal, 2008). Nevertheless, GS has not enabled bulk access to its data, reportedly because their agreements with publishers preclude it (Van Noorden, 2014). Thus, third-party web-scraping software is currently the only practical way to extract more data from GS than permitted by Publish or Perish.

Despite its known errors and limitations, which are consequence of its automated approach to document indexing (Delgado López-Cózar, Robinson-García, & Torres-Salinas, 2014; Jacsó, 2010), GS has been shown to be reliable and to have good coverage of disciplines and languages, especially in the Humanities and Social Sciences, where WoS and Scopus are known to be weak (Chavarro, Ràfols, & Tang, 2018; Mongeon & Paul-Hus, 2016; van Leeuwen, Moed, Tijssen, Visser, & Van Raan, 2001). Analyses of the coverage of GS, WoS, and Scopus across disciplines have compared the numbers of publications indexed or their average citation counts for samples of documents, authors, or journals, finding that GS consistently returned higher numbers of publications and citations (Harzing & Alakangas, 2016; Harzing, 2013; Mingers & Lipitakis, 2010; Prins, Costas, van Leeuwen, & Wouters, 2016). Citation counts from a range of different sources have been shown to correlate positively with GS citation counts at various levels of aggregation (Amara & Landry, 2012; De Groote & Raszewski, 2012; Delgado López-Cózar, Orduna-Malea, & Martín-Martín, 2018; Kousha & Thelwall, 2007; Martín-Martín, Orduna-Malea, & Delgado López-Cózar, 2018; Meho & Yang, 2007; Minasny, Hartemink, McBratney, & Jang, 2013; Moed, Bar-Ilan, & Halevi, 2016; Pauly & Stergiou, 2005; Rahimi & Chandrakumar, 2014; Wildgaard, 2015). See the supplementary materials[1], Delgado López-Cózar et al. (2018); Orduña-Malea, Martín-Martín, Ayllón, and Delgado López-Cózar (2016), and Halevi, Moed, and Bar-Ilan (2017) for discussions of the wider strengths and weaknesses of GS.

A key issue is the ability of GS, WoS, and Scopus to find citations to documents, and the extent to which they index citations that the others cannot find. The results of prior studies are confusing, however, because they have examined different small (with one exception) sets of articles. A summary of the results found in these previous studies is presented in Table 1. For example, the number of citations that are unique to GS varies between 13% and 67%, with the differences probably being due to the study year or the document types or disciplines covered. The only multidisciplinary study (Moed et al., 2016) checked articles in 12 journals from 6 subject areas, which is still a limited set.

The fields previously compared for citation sources (Table 1) are Library and Information Science (5 out of 10 articles analyse case studies about LIS documents/journals/researchers), Medicine (3 papers, analysing oncology, general medicine, and dentistry), Physics (2 articles: general and condensed matter), Chemistry (2 articles: general and inorganic), Computer Science (2 articles: general, and computational linguistics), Biology (2 articles: general, and virology), Social Work, Political Science, and Chinese Studies (1 article each). From this list it is clear that most academic fields have not been analysed for

[1] Supplementary materials available from https://dx.doi.org/10.31235/osf.io/pqr53