Library Hi Tech

Filter bubbles in interdisciplinary research. A Case study on climate and society

Journal:	*Library Hi Tech*
Manuscript ID	LHT-03-2017-0052.R1
Manuscript Type:	Original Article
Keywords:	Online databases, Google Scholar, Web of Science, bibliometrics, open access, Information retrieval

Abstract

Purpose of this paper

In this study, we compare the content of Web of Science and Google Scholar by searching the interdisciplinary field of climate and ancient societies. We aim at analyzing the retrieved documents by open availability, received citations, co-authors and type of publication.

Design/methodolology/approach

We searched the services by a defined set of keyword. Data was retrieved and analyzed using a variety of bibliometric tools such as Publish or Perish, Sci2Tool and Gephi. In order to determine the proportion of open full texts based on the Web of Science result, we relocated the records in Google Scholar, using an off-campus internet connection.

Findings

We found that the top thousand downloadable and analyzable Google Scholar items matched poorly with the items retrieved by Web of Science. Based on this approach (subject-searching), the services appeared complementary rather than similar.

Even though the first search results differ considerably by service, almost each single Web of Science title could be located in Google Scholar. Based on Google Scholar's full text recognition, we found 74 % of Web of Science items openly available and the citation median of these was twice as high as for documents behind paywalls.

Research limitations/implications

Even though our study is a case study, we believe that findings are transferable to other interdisciplinary fields. The share of freely available documents, however, may depend on the investigated field and its culture towards open publishing.

Practical implications

Discovering the literature of interdisciplinary fields puts scholars in a challenging situation and requires a better understanding of the existing infrastructures. We hope our paper contributes to that and can advise the research and library communities.

What is the original/value of paper

In light of an overwhelming and exponentially growing amount of literature, our bibliometric approach is new in a library context.

Introduction

Web of Science (WoS) and Google Scholar (GS) are two of the main tools to identify and access scholarly literature. WoS requires a subscription but offers controlled metatada and advanced search features. GS in turn is freely accessible but has its shortcoming both concerning the use of metadata and searching.

In the last years, a lot has been written about these shortcomings. Even though GS is used extensively by researchers [1], mainly the lack of transparency in regard to coverage and quality is still problematic [e.g. 2, 3]. However, there have been improvements in the algorithm [2], and documents for example are now merged more successfully [4]. While Mikki [5] reported 7.7% duplicates in 2010, four years later Sjögårde [6] reported only 1%. The service seems to be stable over time, although reproduction and verification remains challenging [7, 8]. However, in contrary to the so-called Google filter bubble as coined by Pariser [8] no such effect can be observed in the scholarly context. Based on keyword searching, Yu, Mustapha [9] compared GS results, from IPs located at different geographic locations, finding 90% agreement.

Undoubtedly, the strength of GS compared to WoS lies in its wide content coverage regarding type of publication and field of research. Still, the size of GS is a well-preserved company secret. It is estimated to contain between 100 and 170 million documents [4, 10], which outsizes by far the core collection of WoS, which comprises less than 60 million documents. GS's sovereign position makes the service attractive for both discovery and research assessment exercises [2, 3, 11-13]. Unfortunately, the enormous coverage and applied ranking algorithm, also seem to stop the service from becoming an appropriate tool for scholarly discoveries [2, 14, pp 109].

Open access – literature review

Another considerable asset by GS is the direct hyperlink to the full text wherever available, whether directly through the publishers' web sites, indirectly through library link resolvers and authentication protocols, or open repositories and academic services (e.g. ResearchGate, Academia, or institutional home pages). The share of open publications has been estimated to above 40% by Archambault, Amyot [15]. Similar results are obtained by a recent study regarding highly-cited documents [16]. Jamali and Nabavi [17] and Pitol and De Groote [18] reported the highest shares so far, about 60% and above70% respectively. Open access is advocated widely within academia (even though some voices argue against claiming violation of academic freedom), and accessibility has increased not at least due to funding requirements and imposed governmental and institutional policies. It is however hard to determine its total amount, since open documents are available from various providers, and GS, as the largest aggregator, does not allow massive automated searching. Most of the above mentioned open access studies are therefore case studies.

Whether there exists a citation advantage for open documents has been discussed repeatedly. Arguments against such an advantage are usually related to methodologies and selection procedures of the studies applied [e.g. 19, 20]. Still, the evidence points at a growing citation advantage, and most recent findings [17, 21] report a considerable (50%) higher citation impact for open documents. Whether there is indeed such a citation advantage, is also subject to this article.

Searching by subject – literature review

For GS, only few studies investigate subject searching. These often involve simple and not advanced searches, and their analysis is restricted to the first page of results returned. For example Walters [22] found a higher recall and relevancy for GS results compared to eight other databases for the particular subject field *later-life-migration*. However, this was not the case for more specified and complex searches. Similar results were obtained by Yu, Mustapha [9]. These findings are interesting and worthwhile to investigate further.

Topics related to climate are hot in politics and research, and the scientific output is expected to increase considerably over time. For WoS, the number of documents related to *climate change*, has recently been investigated by Haunschild, Bornmann [23]. The authors retrieved a total of 22000 papers (1980-2014), and reported an exponential growth. They further found that the number of papers related to *adaption, mitigation, risk and vulnerability* were comparatively low, but increasing rapidly. The aspect of *vulnerability* has been studied by Wang, Pan [24], using a stepwise approach to capture the entire literature in WoS (1991-2012). They also report a prominent exponential growth. How a changing climate effects our lives is indeed a major issue in today's research activities.

Inspired by the search methodologies of the mentioned studies, our study investigates the field of *climate impact on societies in the past* and compares the research results from WoS and GS.

This study particularly aims at

- exploring an interdisciplinary field
- designing search strategies and determining overlap of the two services
- analyzing the search results by citations, provided fulltext, title words, author collaborations
- advising the research community

Methodology

We used a quantitative approach to analyze the content of the two citation services Web of Science and Google Scholar.

Subject searching

Defined by a set of keywords, we searched the interdisciplinary field *climate impact on societies in the past* in both services. Boolean operators were applied for WoS, while the advanced search scheme was used for GS. We strived to make the searches act similar and adjusted the expressions slightly, using truncation stars for WoS, confer Expression 1 and 2.

Expression 1 (WOS, see Fig 1):

climat* impact societ* (past or histor* or ancient)

Fig 1. WoS search interface.

Expression 2, same as Expression 1, but omitting truncation stars (GS/PoP, see Fig 2):

climate impact society (past or historical or ancient)

Fig 2. Harzing's Publish or Perish search interface.

The majority of our results is based on these two expression. By applying these expressions however, we learned two lessons:

Lesson 1: The number of results obtained by GS was overwhelming and called for a more careful specification, confer Expression 3.

Lesson 2: The number of results obtained by WoS was not exhaustive and called for a wider formulation including synonyms to increase recall, confer Expression 4.

Based on these lessons we further modified our search results. For GS/PoP we refined the expression and added a geographic region (expression 3) in order to increase precision and thereby decrease the number of recalled documents to a manageable amount. For WoS we added frequently occurring keywords and title words to increase recall (expression 4). These modifications allowed us more correctly to determine similarity of the two the services.

Expression 3 and 4 were defined as follows:

Expression 3 (GS/PoP):

All of the words

<climate human society cultural impact archaeology adaptation resilience vulnerability ancient past>

At least one of the words

<arctic polar "cold regions">

Expression 4 (WoS):

TOPIC: ((societ* (impact* OR adapt* OR collaps* OR resilience* OR vulnerability)) OR (human (impact* OR apapt* OR collaps* OR resilience* OR vulnerability)) OR (*cultur* (impact* OR apapt* OR collaps* OR resilience* OR vulnerability))) AND TOPIC: (*climat*) AND TOPIC: (past OR histor* OR ancient* OR archaeolog* OR holocene OR medieval OR Younger Dryas)

Data retrieval and cleaning

WoS-records were retrieved directly, while GS's top 1000s were retrieved through Harzing's application Publish or Perish (PoP) a free software for analyzing citations [25]. The software has widely been used within academia since its launch in 2006 and is regarded as a complementary service to the commercial tools offered by Clarivate (former Thomson Reuters) and Elsevier. We believe that it is sufficient to look at GS's top 1000 items only, since as a matter of fact no researcher is looking further then the first couple of results pages. Additional data treatment and bibliometric analysis were done in Sci2Tool [26], and analysis on networks were performed in Gephi [27]. Both of these tools are freely available.

Due to the lack of mutual identifiers in the services, we used the author names to determine the degree of similarity. We further made sure that special characters appearing in the author names were treated equally. Furthermore, GS author names were controlled manually to remove items that erroneously were recognized as authors but obviously belonged to different parts of the document. The co-author list returned by GS in general do not exceed more than three authors, hence we know that matches between the services will be incomplete. However, since the aim of our study is only to estimate similarities, we did not clean or enrich the data further (for example by adding missing authors). We also conducted a test where we used the title as a mutual identifier, cleaned the data in LODRefine [28] and merged identical records. We found that both approaches resulted in the same order of overlap, but cleaning the titles was more time consuming. Therefore, we decided to keep the author names as a mutual identifier and as a proxy for estimating the overlap.

In order to determine the proportion of open full texts, we searched GS for either the DOIs or titles provided by WoS from the initial search (Expression 1). As long as a link to a full text was listed, we denoted the status of the document to open access (OA). We did not verify whether the full text was de facto available for each single item. Neither did we examine whether the linked version is a pre-print version or the final publishers' versions nor whether these two differed. In order to avoid paywalled access (through our library SFX link resolvers), we performed the searches off campus.

Automatic sampling was carried out by web scraping, and the following parameters were extracted: Title, Authors, Publication Year, Cited by, format and information on availability (Fig 3). The extracted title was compared with the WoS-title in order to verify similarity.

Fig 3. GS search result, extracted fields highlighted.

Results and discussion

Starting out with searching WoS (Expression 1), we downloaded 639 items. One by one, we then tested whether these items also were indexed by GS. Except two (i.e. 637), all titles could be located. This was an amazingly high recall.

Open access

We found that 468 documents (74%) provided a link to an open full text (Fig 4). The proportion being even higher than reported by Jamali and Nabavi [17] and Martín-Martín, Orduna-Malea [16].

Fig 4. Proportion of open documents (OA) and full text providers (top eight) given by GS.

Figure 4 shows the top eight providers of full text as given by GS. ResearchGate is at the top, followed by Wiley, academia.edu and the American Meteorological Society (ametsoc.com). As the purpose of this study is solely on whether the public has free access or not, we did not distinguish between gold, green, hybrid, legal or illegal access.

Table 1 lists the documents by OA-status. We do not find an obvious increase in open access publishing throughout the decade, but the overall share of OA-documents for this period was as high as 76%.

Table 1. Number and proportion of OA documents and citation median according to GS (2007-2016).

	Documents NON OA	Documents OA	OA %	Citation Median NON OA	Citation Median OA	Fraction of Citation Medians
2007	6	20	77%	25	46	1.8
2008	6	27	82%	27.5	50	1.8
2009	7	26	79%	28	30.5	1.1
2010	7	40	85%	14	33	2.4
2011	11	42	79%	11	21.5	2.0
2012	15	42	74%	10	20	2.0
2013	18	45	71%	7	12	1.7
2014	14	47	77%	5	9	1.8
2015	31	57	65%	2	5	2.5
2016	14	54	79%	2	1	0.5
Totals	**129**	**400**	**76%**	**6**	**13**	**2.2**

We also calculated the citation median for each year and compared the values for OA and NON-OA documents. For all years (except 2016) the citation median was higher for OA documents than for NON-OA documents. In fact, the so-called a-head advantage for the youngest publications is not observed, which might be caused by imposed embargos [17].

For the years shown, the citation median of open documents is 2.2 times the citation median of paywalled documents. It has a maximum in 2010 (2.4), which also correspond to the highest OA-share (85%).

Our findings confirm a strong benefit from open access publishing, and are in agreement with findings by Jamali and Nabavi [17] and the mega study by Archambault, Côté [21].

Subject searching by WoS and GS

Using expression 2 we found 2.5 million items in GS, which outsizes by far the number of documents retrieved by WoS (639), confer Table 2. At the same time, GS does not offer an official API for automatic metadata harvesting and with PoP only a small fraction (1000 documents) is retrievable and analyzable. The rest remains hidden and are therefore questionable. A brief look at the 1000 items shows that titles are highly relevant and confirm GS as a valuable scholarly service.

Table 2. Number of documents and citations in GS and WoS using expression 1 and 2.

	Documents	Citations	Retrieval date
GS estimated total	2590000	NA	31 October 2016
GS retrieved by PoP	1000	310993	31 October 2016
WoS	639	1369	08 November 2016

We observed a pronounced increase of the scholarly literature in the investigated field (Fig 5). This is in accordance to the findings by Haunschild, Bornmann [23] and Wang, Pan [24]. The increase is exponential for WoS during the entire period, while for GS, it decreases during the last 4 years. This is due to GS's algorithm, ranking the most cited documents highest. Since getting cited takes time, the youngest documents most likely won't appear under the top 1000s. Due to differences in size, the citation counts are considerably lower for WoS.

Fig 5. Number of documents by services, WoS and GS top 1000s. 2016 not shown.

For GS, the relative distribution by type of document is shown in Fig 6. Three quarters belong to journal articles, 5% to books, 3% to citing documents. The rest are PDF and HTML documents. The book share was unexpectedly low, given the fact that books in general are more frequently cited [e.g. 16, 17, 29].

Fig 6. Relative distribution by type of document for GS items (all years).

We further estimated the overlap of the two services using the authors' last names and initials. For GS we found that 107 out of 2024 names, about 5%, were identical (Fig 7). Even though the number of authors listed is limited to 3-5 authors for GS, our findings indicate that the overlap is marginal.

Fig 7. Overlap of authors for the two services.

Fig 8 displays the author network of the two services. For GS the network is less crowded and clustered than for WoS. This is mainly due to the fact, that GS lists only 3-5 authors per document. However, we also presume that topics are differently covered and more broadly represented by GS.

Fig 8. Author network for GS top 1000s (left) and WoS (right).

To discover more characteristics of the two services, we extracted the words of the titles and used the stem and stop word analysis by SCi2tool .

Fig 9 shows the top listed title stem words and their co-appearances. The words *Climate*, *Impact* and *Change* are the most frequent words in both of the services. In fact, this is the case for many of the most frequent words. However, they appear in different combinations.

The stem words *China, Environment, Land, Temperatur, Holocen* appear in the top list of WoS but not of GS. On the other hand *Effect, Respons, Affect, Vulner, Forest* appear in the top list of GS but not of WoS. These unique terms might indicate a slightly different subject coverage of the services, shifting towards Social Sciences in GS and towards Natural Sciences in WoS.

Fig 9. Title stem words for GS top 1000s (left) and WoS (right).

We find it problematic that only the top thousand items and not the complete result set from GS is retrievable and analyzable. Our next approach aims therefore at limiting the amount of retrieved results by adding relevant terms from our title and keyword analysis to the search expression (Expression 3). Stepwise, by range of year, we managed to download all retrieved 2249 records (Table 3).

Table 3. Number of retrieved records in GS, based on a revised search expression (Expression 3) and specified by intervals of publishing years.

Arctic	Year interval	Number of documents
GS/PoP	2012-2016	974 (970 downloaded)
	2005-2011	847
	1700-2004	433
GS/POP sum	**1700-2016**	**2254 (2249 downloaded)**

At the expense of journal articles, we found that the book share increased considerably (almost to one-half, Fig 10), resulting in less overlap of the two services. A brief look at the book titles also showed that the returned documents were less relevant, for example 1) *Education, Nature, and Society,* 2) *A Viking Way of Life* and 3) *The Great Perhaps: God as a Question*.

We conclude that carefully specifying the search criteria in GS does not increase precision what suggests that GS uses its metadata insufficiently. In this regard, our findings are in accordance to findings by Walters [22] and Yu, Mustapha [9] .

Fig 10. Type of documents in GS. Search expression refined (Expression 3).

To test the robustness of GS, we also compared results returned by different PCs (work PC and home laptop). The different PCs returned identical results for the top thousand items. Personalization as recorded by e.g. Snipes [30] did not seem to have any effect, and the stated filter bubble [8] couldn’t be detected in Google Scholar, the sub-database of Google. Our findings are in line with findings by Yu, Mustapha [9], where similarity of search results was reported to above 90%, and being independent on geographic region.

Using Expression 1 for searching WoS returned 639 results only, as shown in Table 2. We understood that this number was far from exhaustive and that the expression needed revision. We therefore added frequently occurring keywords and title words to increase recall (Expression 4).

The improved search expression returned 6643 results, about ten times the initial result. The number of similar authors for the services increased to 787 (Fig 11), which corresponds to 4 % overlap compared to 5% before. These results show that subject indexing in WoS is insufficient. The service only superficially indexes its documents. It seems to be up to the user to carefully design the searches and add all possible synonyms. Consequently, the probability to miss relevant documents is high.

Fig 11. Overlap of author names in the two services with a modified search for WoS (Expression 4).

Conclusion and final remarks

We compared the search results of two of the main tools to access scholarly literature, WoS and GS and investigated the interdisciplinary field *climate impact on ancient societies* which covers the humanities, social sciences and natural sciences.We found that each single WoS title (except two) could be located in GS. This confirms GS sovereignty as a source for scholarly literature. According to GS full text recognition, we found 74% of the documents openly available either directly on the publishers’ websites, or indirectly in repositories or in other ways. The citation median of open documents is more than twice the median of paywalled documents. Obviously, full text links provided by GS has been essential for the transition towards open publishing, and our findings challenge the traditional subscription-based publishing model.

Starting out with a simple search expression, we estimated the overlap between the services to 5%, considering GS top 1000 items only. This comparison was based on the authors' last name and initials. The overlap increased to 40% when the search expression was enhanced for WoS. A carefully specified search for GS on the other hand, limited the number of returned documents, but unfortunately, did not increase precision and relevancy. These findings indicate that the use of metadata is insufficient and conflicts with the scholars' need to perform sound literature reviews. However, our findings also indicate that GS is capable of locating relevant documents without carefully constructing advanced searches. We learned further that the two evaluated services function differently in their logic. This is something to take into account for future searching and library teaching.

The network analysis revealed that subjects are slightly differently covered by the services. As expected, natural science related documents were more prevalent in WoS, while social science related documents were more prevalent in GS.

Applying frequent title words and keywords to enhance the search expression for WoS proved useful, and the overlap of the two services increased from 5% to 40 % (still keeping in mind that only GS top 1000 items are considered). It also proved that the service only shallowly indexes its content.

We conclude that neither WoS nor GS can be used as stand-alone service to discover the scholarly literature of the investigated field. The services returned complementary rather than similar results. They may be interpreted as almost decoupled filter bubbles. Our findings also indicate that the recalled documents only reflect a fraction of the total amount of the entire scholarly content. In order to discover the remaining literature, a follow-up study may investigate additional sources such as library discovery tools and discipline specific databases.

In light of an overwhelming and exponentially growing amount of literature, our bibliometric approach is new in a library context and much needed by the academic community. In particular, discovering the literature of interdisciplinary fields puts scholars in a challenging situation. First, terminologies used by the disciplines differ, second, the information and communication systems are separated and third, researchers are torn between different scholarly cultures making it hard to bridge the gap between them. A call for increased interdisciplinary research requires a better understanding and an adaption of the research infrastructure [31, 32].

References

1. Van Noorden R. Scientists and the social network. Nature. 2014;512(7513):126-9. doi: 10.1038/512126a.
2. Jacsó P. Metadata mega mess in Google Scholar. Online Information Review. 2010;34(1):175-91. doi: 10.1108/14684521011024191.
3. Prins AA, Costas R, van Leeuwen TN, Wouters PF. Using Google Scholar in research evaluation of humanities and social science programs: A comparison with Web of Science data. Research Evaluation [Internet]. 2016:[rvv049 p.].
4. Orduña-Malea E, Ayllón JM, Martín-Martín A, López-Cózar ED. About the size of Google Scholar: playing the numbers. arXiv:14076239. 2014.
5. Mikki S. Comparing Google Scholar and ISI Web of Science for Earth Sciences. Scientometrics. 2010;82(2):321-31. doi: 10.1007/s11192-009-0038-6.
6. Sjögårde P. Jämförelse mellan Google Scholar, Scopus och Web of Science – En fallstudie av en Unit of Assessment i RAE2012. Stockholm: KTH; 2014. Available from: https://www.kth.se/polopoly_fs/1.508325!/J%C3%A4mf%C3%B6relse%20mellan%20Web%20of%20Sceince,%20Scopus%20och%20Google%20Scholar.pdf.
7. Bramer WM. Variation in number of hits for complex searches in Google Scholar. Journal of the Medical Library Association. 2016;104(2):143-5. doi: 10.3163/1536-5050.104.2.009.
8. Pariser E. The filter bubble: What the Internet is hiding from you: Penguin UK; 2011.
9. Yu K, Mustapha N, Oozeer N. Google Scholar's Filter Bubble. In: Esposito A, editor. Research 20 and the Impact of Digital Technologies on Scholarly Inquiry: IGI Global; 2017. p. 211-29. doi: 10.4018/978-1-5225-0830-4.ch011.
10. Khabsa M, Giles CL. The number of scholarly documents on the public web. PLoS One. 2014;9(5):e93949. doi: 10.1371/journal.pone.0093949.
11. Harzing A-W. A preliminary test of Google Scholar as a source for citation data: a longitudinal study of Nobel prize winners. Scientometrics. 2013;94(3):1057-75. doi: 10.1007/s11192-012-0777-7.
12. Harzing A-W. A longitudinal study of Google Scholar coverage between 2012 and 2013. Scientometrics. 2014;98(1):565-75. doi: 10.1007/s11192-013-0975-y.
13. LSE Public Policy Group. Maximizing The Impacts Of Your Research: A Handbook For Social Scientists 2011. Available from: http://blogs.lse.ac.uk/impactofsocialsciences/2011/04/14/maximizing-the-impacts-of-your-research-a-handbook-for-social-scientists-now-available-to-download-as-a-pdf/.
14. Ortega JL. Academic search engines: A quantitative outlook: Elsevier; 2014.
15. Archambault E, Amyot D, Deschamps P, Nicol A, Rebout L, Roberge G. Proportion of Open Access Peer-Reviewed Papers at the European and World Levels—2004-2011 2013. Available from: http://www.science-metrix.com/pdf/SM_EC_OA_Availability_2004-2011.pdf.
16. Martín-Martín A, Orduna-Malea E, Ayllón JM, Delgado López-Cózar E. A two-sided academic landscape: portrait of highly-cited documents in Google Scholar (1950-2013). Revista Española de Documentación Científica. 2016;Preprint. doi: 10.3989/redc.2016.4.1405.

17. Jamali HR, Nabavi M. Open access and sources of full-text articles in Google Scholar in different subject fields. Scientometrics. 2015;105(3):1635-51. doi: 10.1007/s11192-015-1642-2.

18. Pitol SP, De Groote SL. Google Scholar versions: do more versions of an article mean greater impact? Library Hi Tech. 2014;32(4):594-611. doi: 10.1108/LHT-05-2014-0039.

19. Hersh G, Plume A. Citation metrics and open access: what do we know? : Elsevier Connect; 2016 [09.12.16]. Available from: https://www.elsevier.com/connect/citation-metrics-and-open-access-what-do-we-know.

20. Hua F, Sun HY, Walsh T, Worthington H, Glenny AM. Open access to journal articles in dentistry: Prevalence and citation impact. J Dent. 2016;47:41-8. doi: 10.1016/j.jdent.2016.02.005.

21. Archambault E, Côté G, Struck B, Voorons M. Research impact of paywalled versus open access papers. Science-Metrix and 1science [Internet]. 2016. Available from: http://www.1science.com/oanumbr.html.

22. Walters WH. Comparative Recall and Precision of Simple and Expert Searches in Google Scholar and Eight Other Databases. Portal-Libraries and the Academy. 2011;11(4):971-1006. doi: 10.1353/pla.2011.0042

23. Haunschild R, Bornmann L, Marx W. Climate change research in view of bibliometrics. PLoS One. 2016;11(7):e0160393. doi: 10.1371/journal.pone.0160393.

24. Wang B, Pan SY, Ke RY, Wang K, Wei YM. An overview of climate change vulnerability: a bibliometric analysis based on Web of Science database. Natural Hazards. 2014;74(3):1649-66. doi: 10.1007/s11069-014-1260-y.

25. Harzing A-W. Publish or Perish [cited 2011 July 18th]. Available from: http://www.harzing.com.

26. Sci2 Team. Science of Science (Sci2) Tool. Indiana University and SciTech Strategies; 2009.

27. Bastian M, Heymann S, Jacomy M. Gephi: an open source software for exploring and manipulating networks. ICWSM [Internet]. 2009; 8:[361-2 pp.]. Available from: http://www.aaai.org/ocs/index.php/ICWSM/09/paper/viewFile/154/1009/.

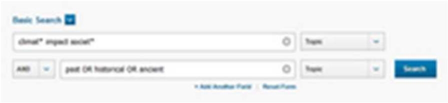
28. Sparkica. LodRefine 1.0.7.1. Available from: http://openrefine.org/.

29. Nederhof AJ, Van Raan AF. A bibliometric analysis of six economics research groups: A comparison with peer review. Research Policy. 1993;22(4):353-68. doi: 10.1016/0048-7333(93)90005-3.

30. Snipes S. 2012. [27.10.2016]. Available from: https://www.qdigitalstudio.com/library/reasons-your-google-search-results-are-different-than-mine.

31. Gullbekk E. Apt information literacy? A case of interdisciplinary scholarly communication. Journal of Documentation. 2016;72(4):716-36. doi: 10.1108/JDOC-08-2015-0101.

32. Brister E. Disciplinary capture and epistemological obstacles to interdisciplinary research: lessons from central African conservation disputes. Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences. 2016;56:82-91. doi: 10.1016/j.shpsc.2015.11.001.

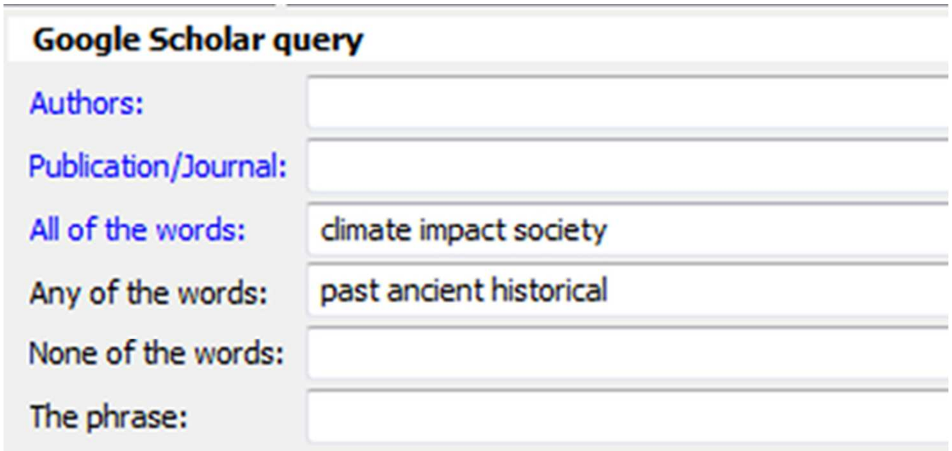


climat* impact social* Topic

AND past OR historical OR ancient Topic

Fig 1: WoS search interface.

18x4mm (300 x 300 DPI)



Google Scholar query

Authors:	
Publication/Journal:	
All of the words:	climate impact society
Any of the words:	past ancient historical
None of the words:	
The phrase:	

Fig 2: Harzing’s Publish or Perish search interface.

39x18mm (300 x 300 DPI)

[PDF] The effects and consequences of very large explosive volcanic eruptions

[PDF] royalsocietypublishing.org

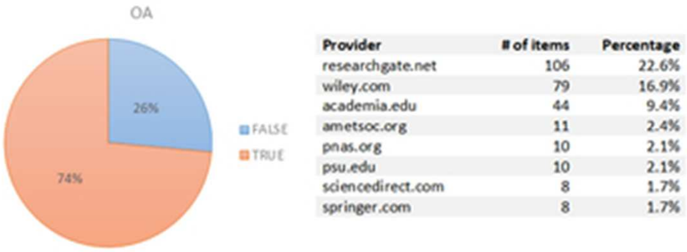
[illegible] Transactions of the Royal Society A [illegible] 2006 [illegible]

[illegible]

Related articles [illegible]

Fig 3: GS search result, extracted fields highlighted.

14x2mm (300 x 300 DPI)



Provider	# of items	Percentage
researchgate.net	106	22.6%
wiley.com	79	16.9%
academia.edu	44	9.4%
ametsoc.org	11	2.4%
pnas.org	10	2.1%
psu.edu	10	2.1%
sciencedirect.com	8	1.7%
springer.com	8	1.7%

Fig 4: Proportion of open documents, full text providers (top eight) given by GS.

30x11mm (300 x 300 DPI)

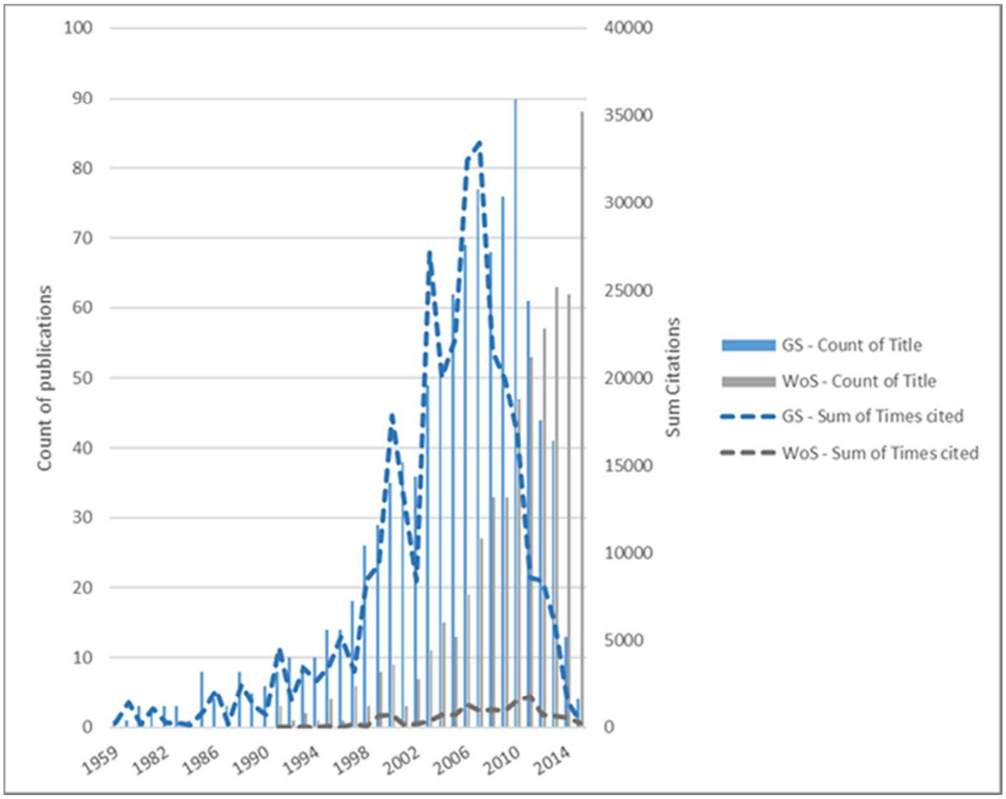


Fig 5: Number of documents by services, WoS and GS top 1000s. 2016 not shown.

66x52mm (300 x 300 DPI)

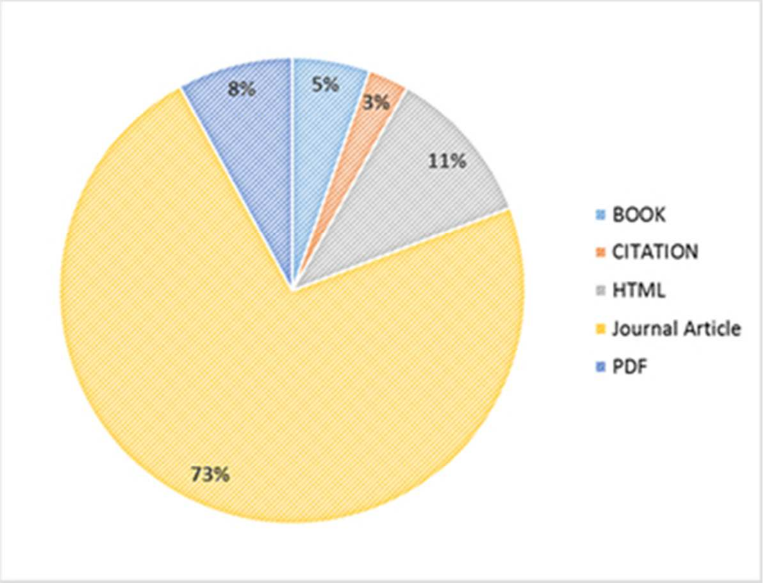


Fig 6: Relative distribution by type of document for GS items (all years).

32x24mm (300 x 300 DPI)

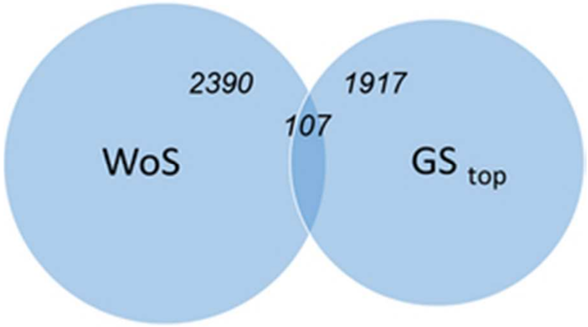


Fig 7: Overlap of authors for the two services.

24x14mm (300 x 300 DPI)

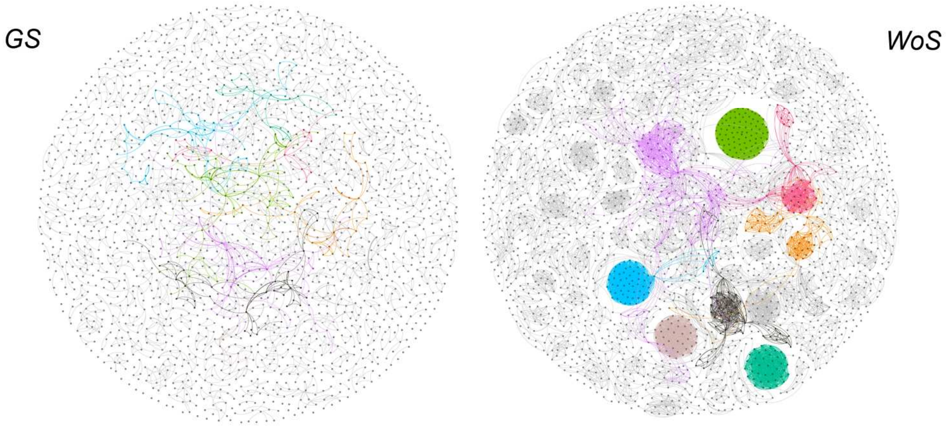


Fig 8: Author network GS top 1000s (left) and WoS (right).

190x81mm (300 x 300 DPI)

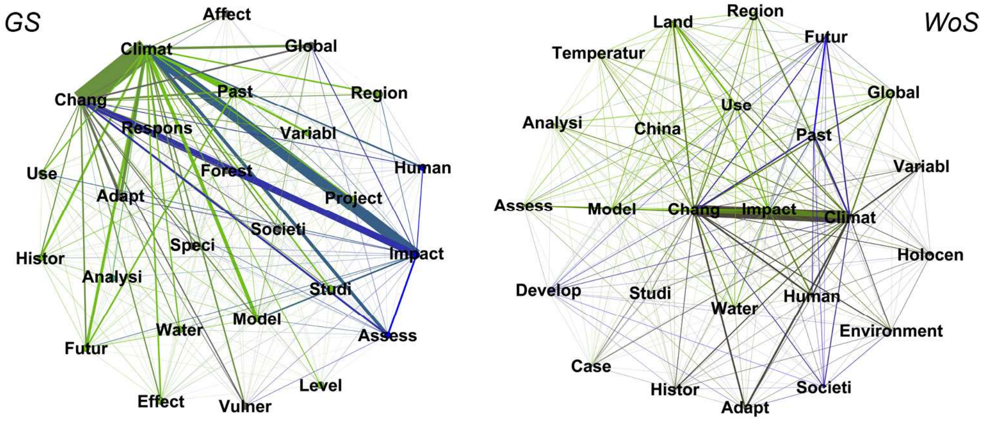


Fig 9: Title stem words GS top 1000s (left) and WoS (right).

190x79mm (300 x 300 DPI)

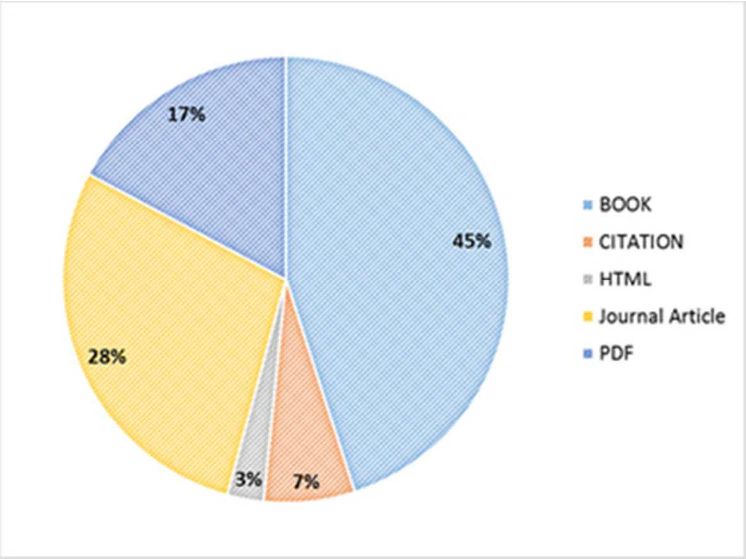


Fig 10: Type of documents in GS. Search expression refined.

31x23mm (300 x 300 DPI)

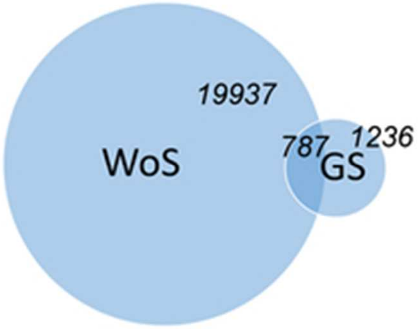


Fig 11: Overlap of authors in the two services with a modified search for WoS.

24x14mm (300 x 300 DPI)