



UNIVERSIDAD
DE GRANADA

SUHF



UPPSALA
UNIVERSITET

Google Scholar's citation graph: comprehensive, global... and inaccessible

Alberto Martín-Martín, Emilio Delgado López-Cózar

Open Citations seminar, May 23rd, 2018, Uppsala, Sweden

TEAM



EMILIO DELGADO LÓPEZ-CÓZAR
PROFESSOR
@ UNIVERSIDAD DE GRANADA



ENRIQUE ORDUÑA-MALEA
ASSISTANT PROFESSOR
@ UNIVERSIDAD POLITÉCNICA
DE VALENCIA



ALBERTO MARTÍN-MARTÍN
PHD STUDENT
@ UNIVERSIDAD DE GRANADA



STRUCTURE OF THE TALK



CONTEXT

How we got to this point



STRENGTHS

of data available in Google Scholar



WEAKNESSES

of data available in Google Scholar



PERSONAL EXPERIENCES

getting data from Google Scholar to generate data products





CONTEXT

HOW WE GOT TO THIS POINT

SPAIN (1996-2010)

IN-RECS Buscar

ÍNDICE DE IMPACTO
REVISTAS ESPAÑOLAS DE CIENCIAS SOCIALES
DOCUMENTACIÓN

Ayuda | Estadísticas | Revistas fuente

Revistas														Artículos					Autores			Instituciones		
Impacto por años														Impacto acumulativo										
2010	2009	2008	2007	2006	2005	2004	2003	2002	2001	2000	1999	1998	1997	1996	Histórico	2000-2009	2005-2009							

ÍNDICE DE IMPACTO: 2010

Población de revistas: 33

CUARTIL	POSICIÓN	TÍTULO DE LA REVISTA	ÍNDICE IMPACTO 2010	TOTAL ARTÍCULOS	TOTAL CITAS	CITAS NACIONALES	CITAS INTERNACIONALES
1º	1	El Profesional de la Información	0.578	83	48	37	11
	2	Cybermetrics: International Journal of Scientometrics, Informetrics and Bibliometrics	0.500	14	7	1	6
2º	2	Revista Española de Documentación Científica	0.500	56	28	22	6
	3	BiD: Textos Universitarios de Biblioteconomía i Documentació	0.220	59	13	13	0
	4	Anales de Documentación	0.148	54	8	6	2
3º	5	Anuario Hipertext.net	0.118	17	2	2	0
	5	Papeles Médicos. Revista de la Sociedad Española de Documentación Médica	0.118	17	2	2	0
	6	Documentación de las Ciencias de la Información	0.093	43	4	4	0
4º	7	Item. Revista de Biblioteconomía i Documentació	0.087	46	4	4	0
	8	Anuario ThinkEPI	0.075	133	10	6	4
	9	Revista General de Información y Documentación	0.030	67	2	2	0
	10	Museo. Revista de la Asociación Profesional de Museólogos de España	0.013	77	1	1	0
	11	Boletín de la ANABAD	0.010	206	2	2	0
	12	Educación y Biblioteca. Revista Mensual de Documentación y Recursos Didácticos	0.009	350	3	3	0
	13	Aedom. Boletín de la Asociación Española de Documentación Musical	0.000	13	0	0	0
	13	Bilduma: Revista del Servicio de Archivo del Ayuntamiento de Errenteria	0.000	9	0	0	0
	13	Boletín de la Asociación Andaluza de Bibliotecarios	0.000	44	0	0	0
	13	Cartas Diferentes: revista canaria de patrimonio documental	0.000	26	0	0	0
	13	Cuadernos de Documentación Multimedia	0.000	28	0	0	0
	13	Cultura escrita y sociedad	0.000	60	0	0	0
	13	Elucidario: Seminario bio-bibliográfico Manuel Caballero Venzalá	0.000	130	0	0	0
	13	Ibersid. Revista de Sistemas de Información y Documentación	0.000	116	0	0	0
	13	Lligall. Revista Catalana d'Arxivística	0.000	24	0	0	0
13	Ocnos: revista de estudios sobre lectura	0.000	21	0	0	0	
13	Pecia complutense	0.000	25	0	0	0	
13	PH. Boletín del Instituto Andaluz del Patrimonio Histórico	0.000	141	0	0	0	
13	Red Iris. Boletín de la Red Nacional de I+D	0.000	56	0	0	0	
13	Revista d'arxius	0.000	20	0	0	0	
13	Revista de Museología	0.000	126	0	0	0	
13	Scire. Representación y Organización del Conocimiento	0.000	47	0	0	0	

<http://ec3.ugr.es/in-recs/>



CITATION INDEXES



WEB OF SCIENCE™



Scopus®

- Selective coverage based on source selection
- Commercial (subscription-based). License to access to data in bulk separate from license to web application



- Inclusive coverage based on parsing webpages
- Non-commercial service offered by Google. Free to access. Doesn't offer options to access data in bulk (agreements with publishers preclude it)



THE NEED OF OPEN CITATIONS



In this open-access age, it is a scandal that reference lists from journal articles [...] are not readily and freely available for use by all scholars.



DAVID SHOTTON
FOUNDER OF OCC
[source](#)



The citation graph is one of humankind's most important intellectual achievements



DARIO
TARABORELLI
FOUNDER OF I4OC
[source](#)



[I]n order to guarantee full transparency and reproducibility of scientometric analyses, these analyses need to be based on open data sources



ISSI
[source](#)

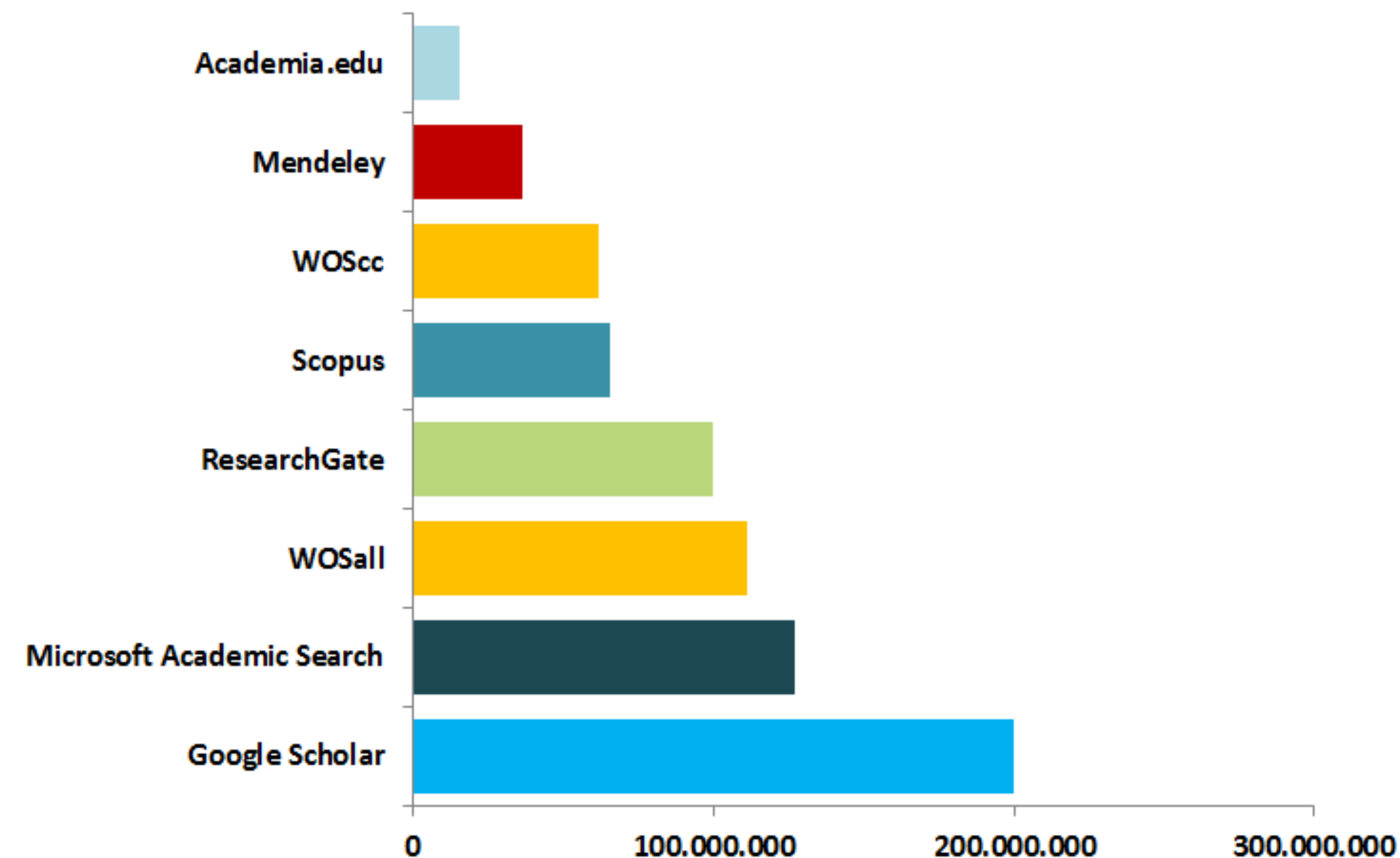


STRENGTHS

OF DATA AVAILABLE IN GOOGLE SCHOLAR

OVERALL SIZE

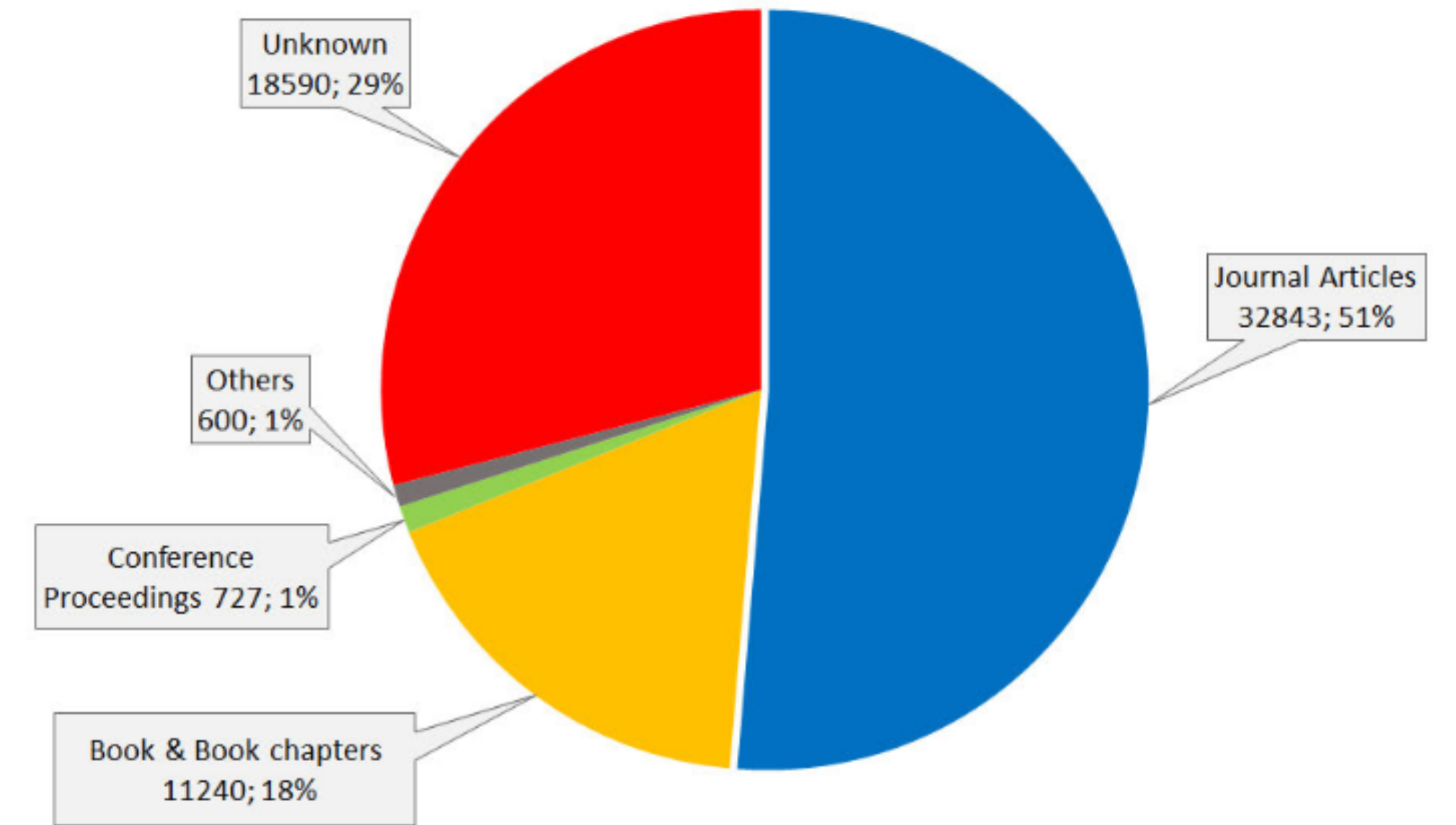
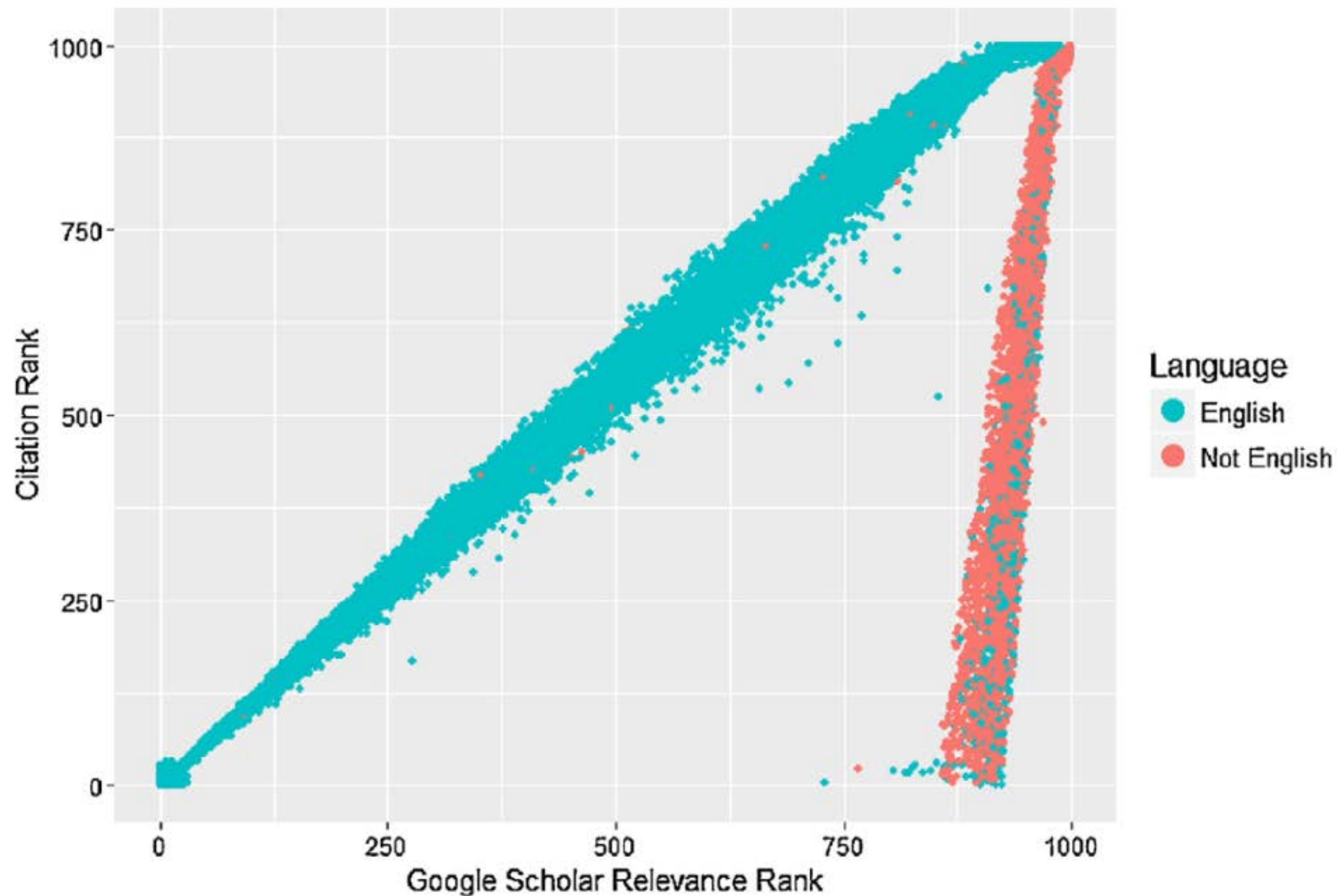
- Khabsa & Giles (2014): around 100 million documents (only in English)
- Orduna-Malea *et al.* (2015): 130-180 million documents (no language restrictions)
- Roughly 2-3 times the size of Web of Science and Scopus. There are disciplinary differences as we'll see later on



- Khabsa, M., & Giles, C. L. (2014). The number of scholarly documents on the public web. *PloS one*, 9(5), e93949. <https://doi.org/10.1371/journal.pone.0093949>
- Orduna-Malea, E., Ayllón, J. M., Martín-Martín, A., & Delgado López-Cózar, E. (2015). Methods for estimating the size of Google Scholar. *Scientometrics*, 104(3), 931-949. <https://doi.org/10.1007/s11192-015-1614-6>



DOCUMENT COVERAGE (1)



- For a sample of 64,000 highly-cited documents according to Google Scholar, 49% of these documents were not covered by Web of Science

- Martin-Martin, A., Orduna-Malea, E., Harzing, A. W., & Delgado López-Cózar, E. (2017). Can we use Google Scholar to identify highly-cited documents?. *Journal of Informetrics*, 11(1), 152-163. <https://doi.org/10.1016/j.joi.2016.11.008>
- Martín-Martín, A., Orduna-Malea, E., Ayllón, J. M., & Delgado López-Cózar, E. (2016). A two-sided academic landscape: snapshot of highly-cited documents in Google Scholar (1950-2013). *Revista Española de Documentación Científica*, 39(4), 1. <https://doi.org/10.3989/redc.2016.4.1405>



DOCUMENT COVERAGE (2)

- “Classic Papers”: Highly cited documents published in 2006 according to Google Scholar (released in June 2017)
- 252 unique subcategories, 8 broad categories covering all areas of knowledge
- 10 most cited documents in each subcategory. At least 20 citations per paper. Total number of articles: 2,515 (one category had only 5 documents)

Category	Number of documents	Not found in WoS Not found in Scopus	
		(%)	(%)
Humanities, Literature & Arts	245	28.2	17.1
Social Sciences	510	17.5	8.6
Engineering & Computer Science	570	11.6	2.5
Business, Economics & Management	150	6.0	2.7
Health & Medical Sciences	680	2.8	0.3
Physics & Mathematics	230	2.2	1.7
Life Sciences & Earth Sciences	380	0.5	0.5
Chemical & Material Sciences	170	0	0

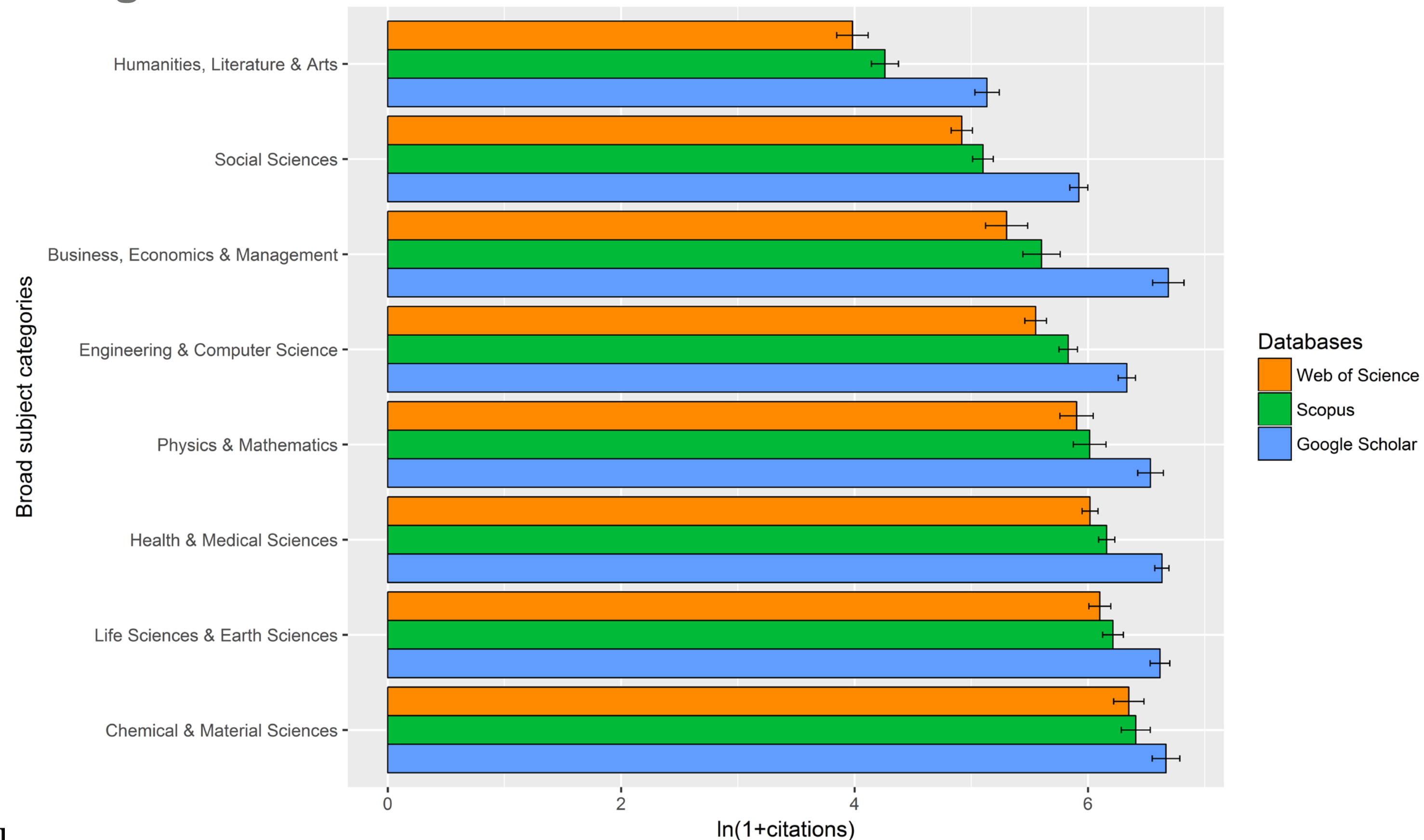
Martín-Martín, A., Orduna-Malea, E., & López-Cózar, E. D. (2018, April 23). Coverage of highly-cited documents in Google Scholar, Web of Science, and Scopus: a multidisciplinary comparison.

<http://doi.org/10.17605/OSF.IO/HCX27>



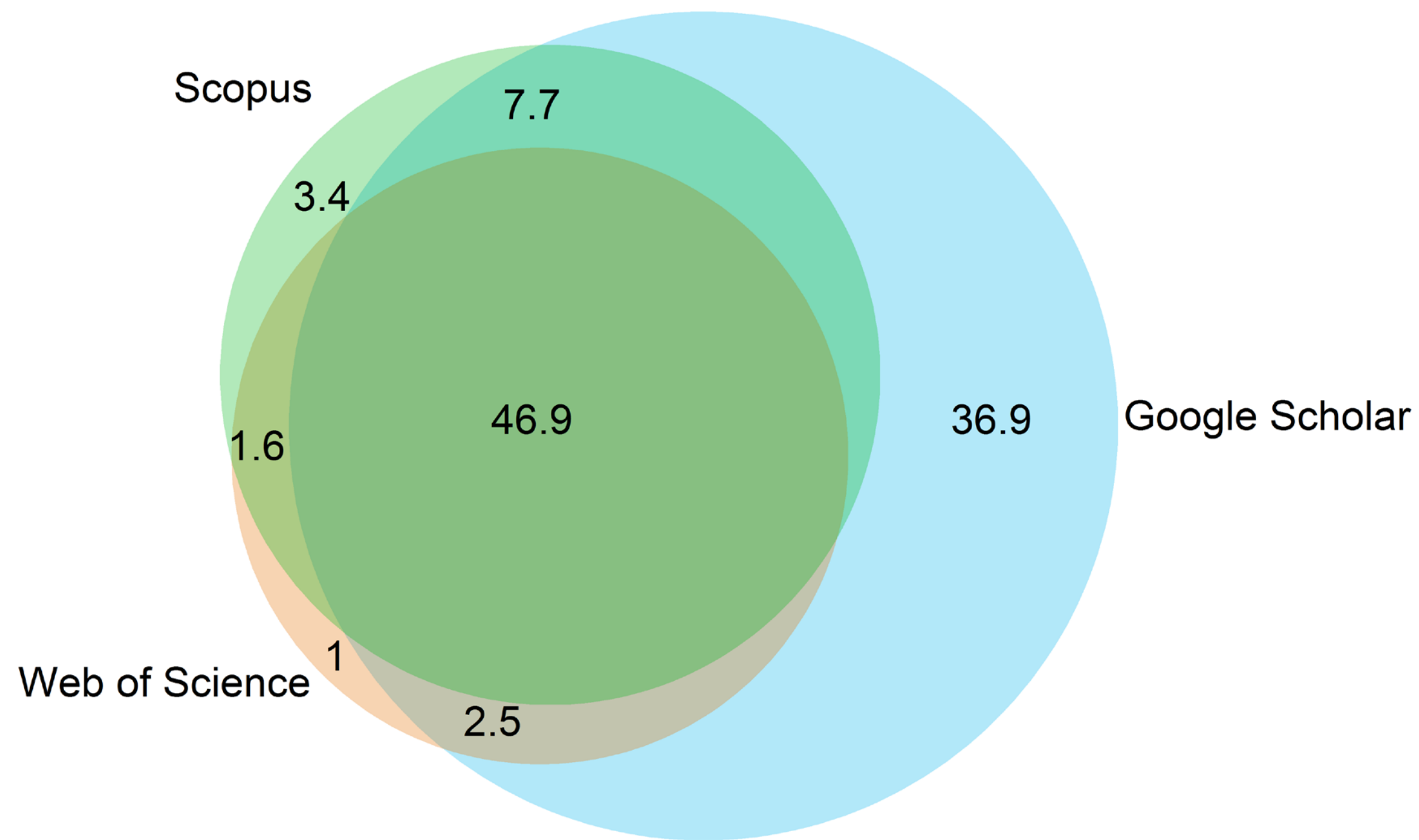
CITATIONS FOUND (1)

Average log-transformed citation counts of highly-cited documents according to Google Scholar published in 2006, based on data from Google Scholar, Web of Science, and Scopus, by broad subject categories



CITATIONS FOUND (2)

We extracted the list documents that cite these highly-cited documents from GS (custom script), WoS (web export), and Scopus (web export)



2.30
M

Total number of citations to 2,299 highly-cited documents from 2006 covered by GS, WoS, and Scopus

95%

Of all citations found by WoS (1.27 M) are also found by Google Scholar

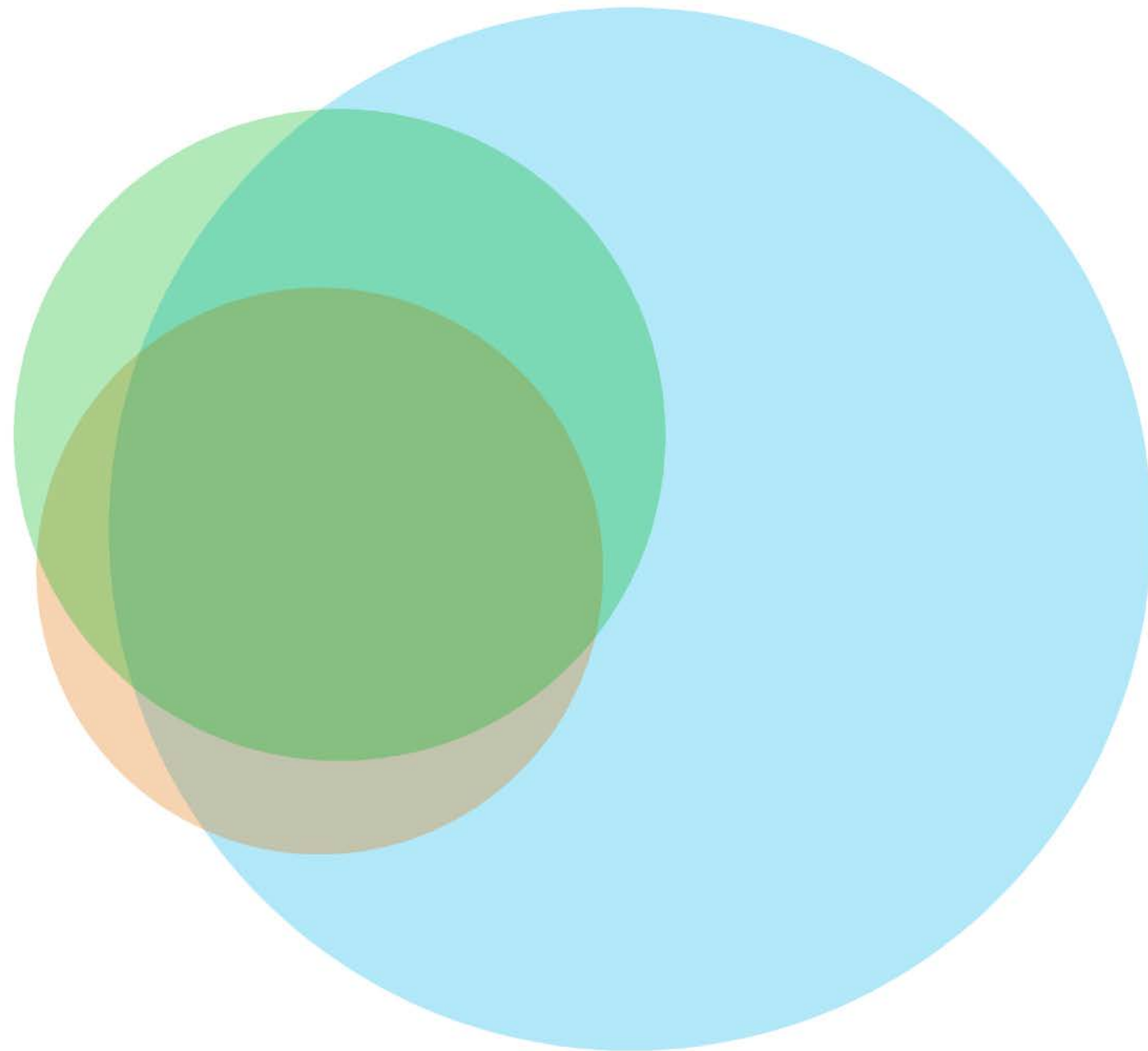
91%

Of all citations found by Scopus (1.47) are also found by Google Scholar

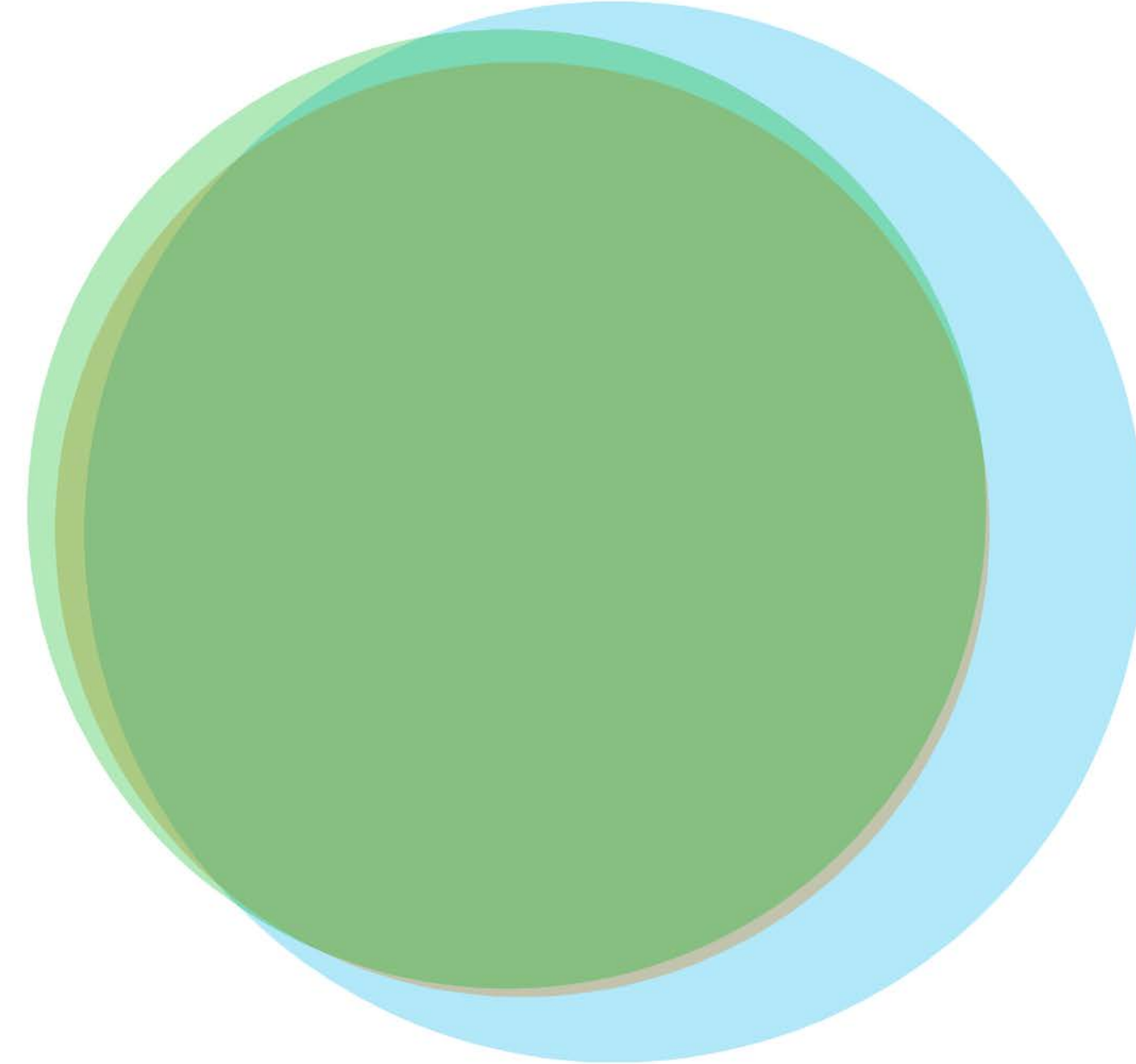


CITATIONS FOUND (3)

HUMANITIES, LITERATURE & ARTS



CHEMICAL & MATERIAL SCIENCES

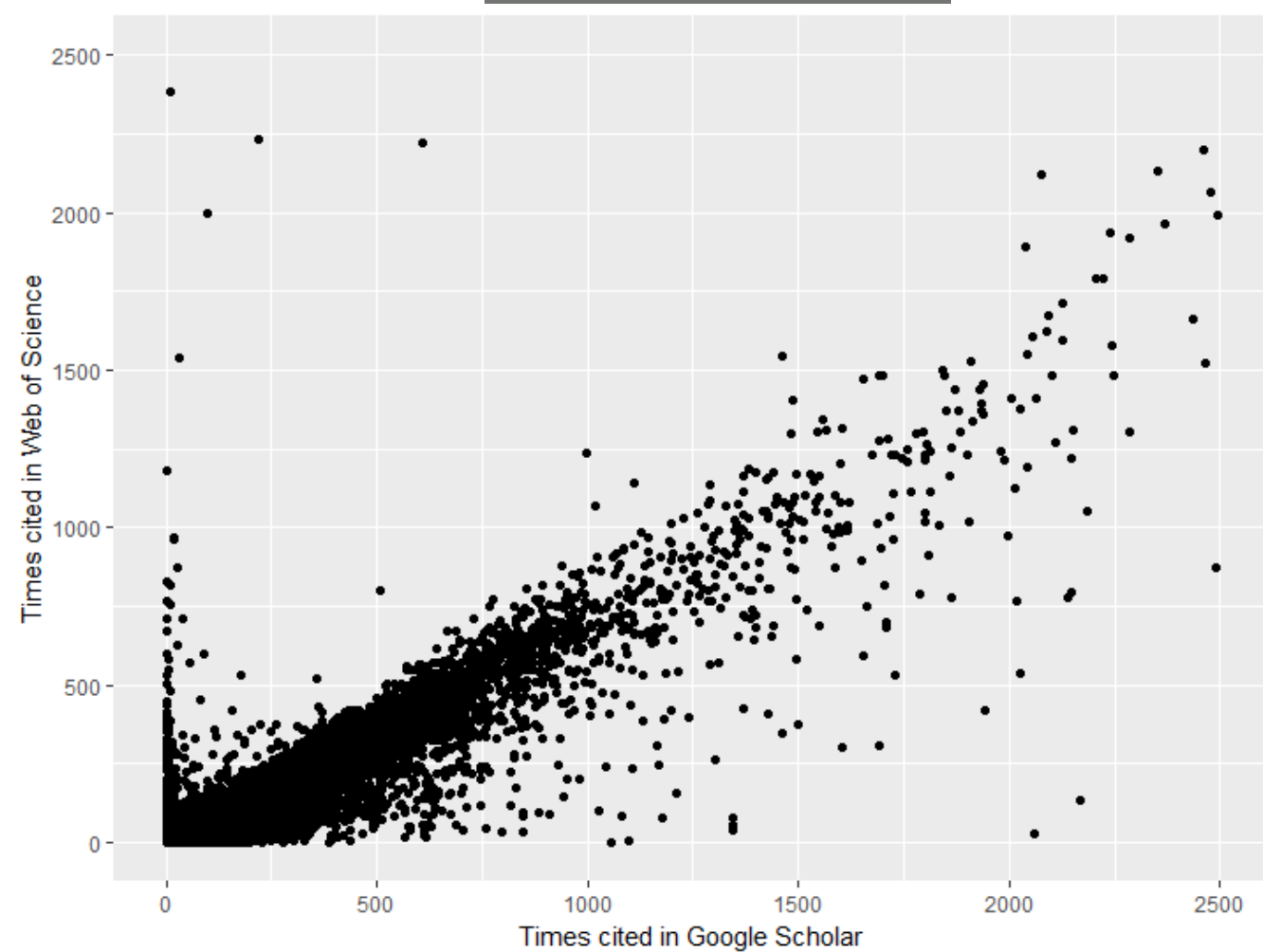


What sources / document types does GS cover that WoS and Scopus do not?

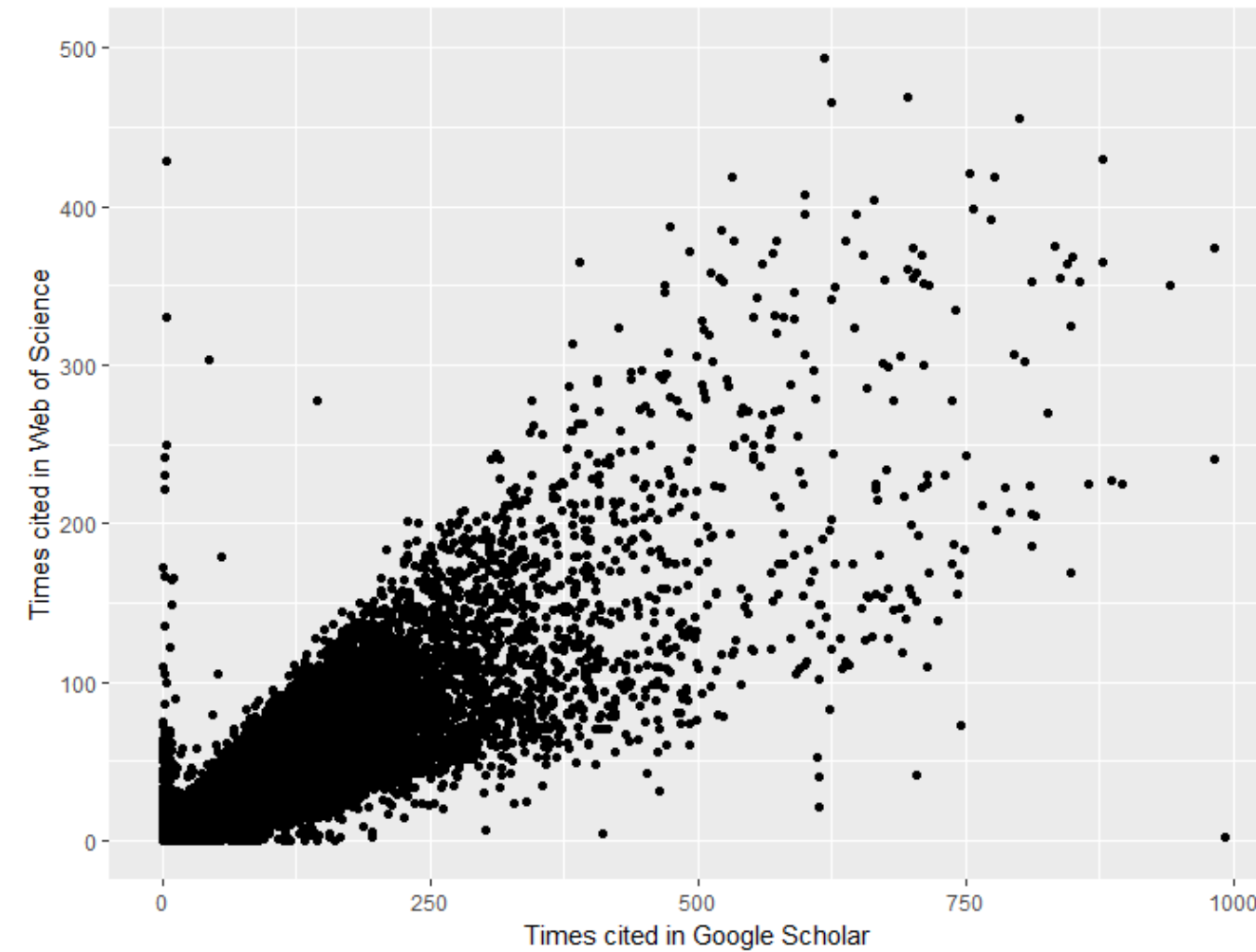
CITATIONS FOUND (4)

- Analysis of articles or reviews with a DOI published in 2009 covered by Web of Science and Google Scholar (~1 million documents)

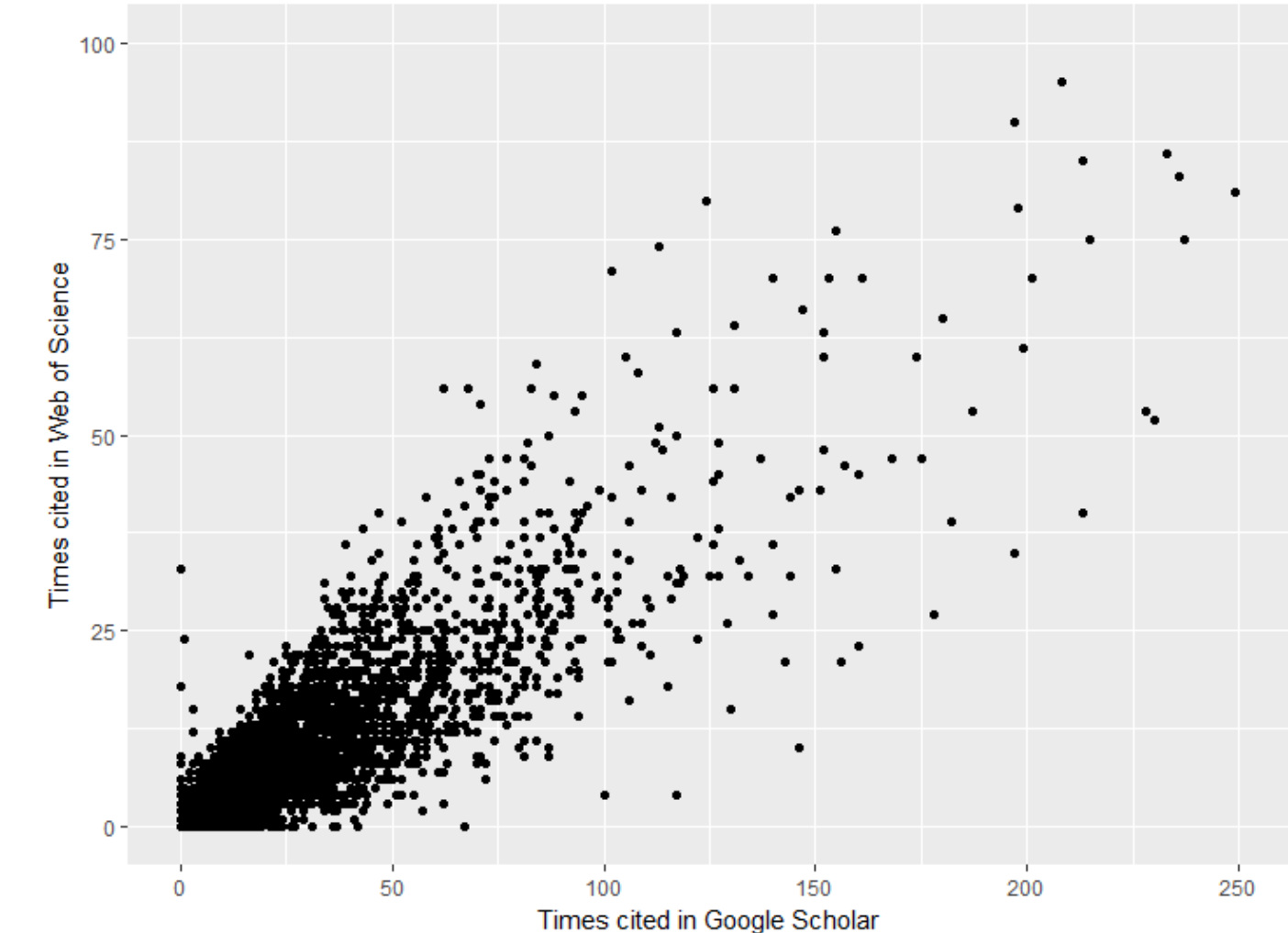
Sciences



Social Sciences



Arts & Humanities



Citation Index	N	spearman.cor	p.value	prop.cited.gs	prop.cited.wos	ratio of gs_cit to wos_cit (avg)
Sciences	863801	0,94	0,00	0,97	0,95	1,68
Social Sciences	109232	0,90	0,00	0,97	0,94	2,58
Art & Humanities	13487	0,83	0,00	0,84	0,69	2,52

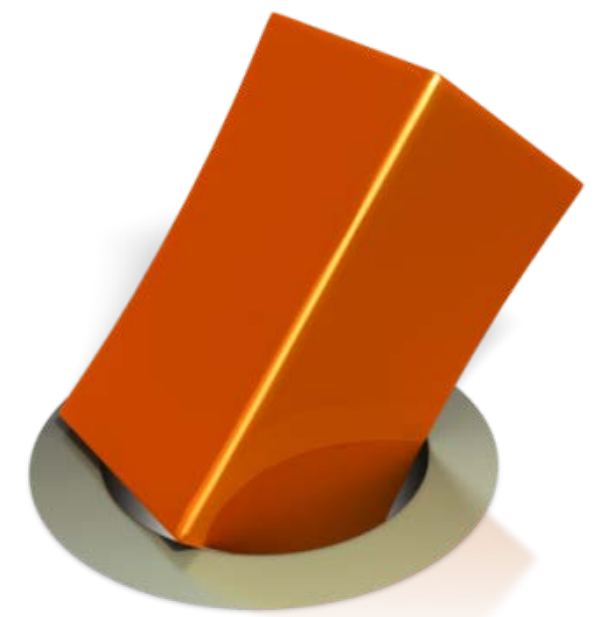


WEAKNESSES

OF DATA AVAILABLE IN GOOGLE SCHOLAR

LIMITATIONS (1)

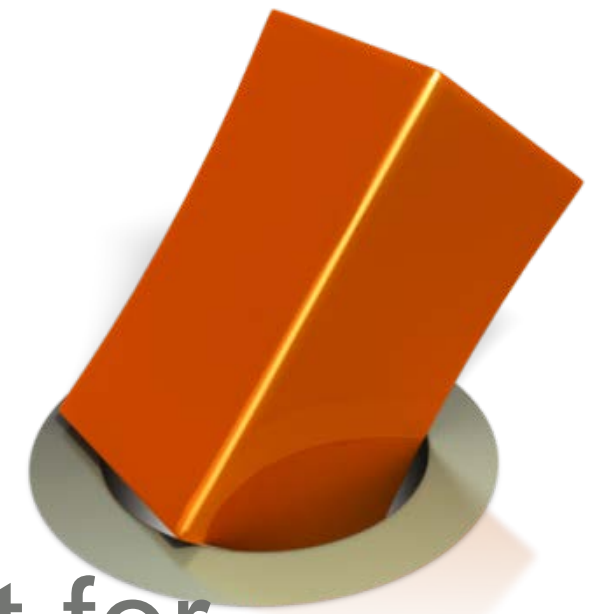
CONSEQUENCE OF TRYING TO USE A TOOL FOR A PURPOSE IT WASN'T DESIGNED FOR



- No support for data reuse: **no API** available. All data extraction has to be made through web scraping
 - Agreements with publishers preclude them from releasing the data
 - Tight security measures to avoid massive data collection (CAPTCHAs)
- **Persistent identifiers** (DOIs, ORCIDs...) are not available to the public (although they use DOIs internally)
- **Only 1,000 results** can be displayed for any given query
- Inability to fix **individual errors**. Very small team of people working on GS. Everything is automated.
- **Incorrect assignment of documents** to researcher profiles (GSC)

LIMITATIONS (2)

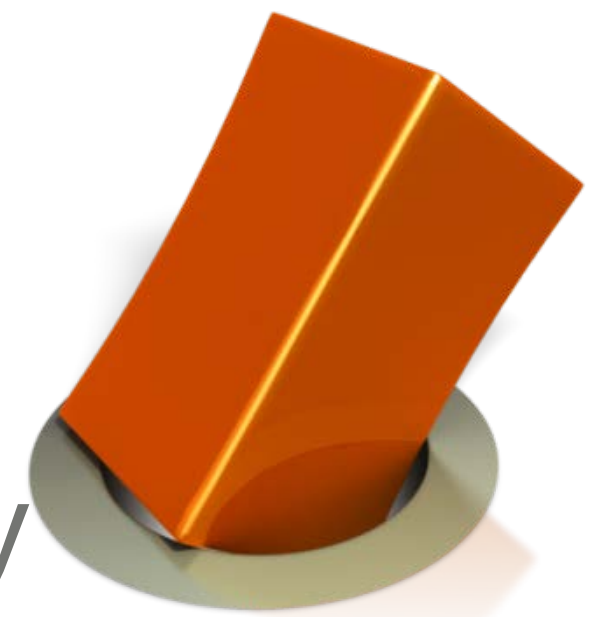
CONSEQUENCE OF TRYING TO USE A TOOL FOR A PURPOSE IT WASN'T DESIGNED FOR



- **Incomplete or erroneous basic metadata**, with little or no support for categorical variables at the document level:
 - incomplete lists of authors!!
 - truncated journal names!!
 - no document types
 - no author affiliations (only at author-level in GSC)
 - no subject classifications
 - Forget about funding acknowledgements...
- Undetected **duplicates**
- Open to **manipulation**
 - Delgado López-Cózar, E., Robinson-García, N., & Torres-Salinas, D. (2014). The Google Scholar experiment: How to index false papers and manipulate bibliometric indicators. *Journal of the Association for Information Science and Technology*, 65(3), 446-454.
<https://doi.org/10.1002/asi.23056>
- **Cited references** of documents are not available

LIMITATIONS (3)

IS THERE A WAY TO OVERCOME THEM...



... that doesn't involve paying mind-numbingly high amounts of money to some multinational?

- **Complementing Google Scholar metadata** with data from other freely accessible sources:
 - Going to the source: Metadata in **publisher websites or repositories**
 - **CrossRef Metadata API** (for everything with a DOI)
 - Complete basic metadata.
 - Cited references available for over 51% of their records (so far Springer, Wiley, and some smaller publishers have agreed to make them public). <https://i4oc.org/>
 - Author affiliations available from some publishers
 - **Digging deeper into Google Scholar:** they have more metadata, but it's expensive to extract it (time-wise).



PERSONAL EXPERIENCES

GETTING DATA FROM GOOGLE SCHOLAR TO GENERATE
DATA PRODUCTS

SOFTWARE THAT HANDLES GOOGLE SCHOLAR DATA

- **Publish or Perish**, by Anne-Wil Harzing: <https://harzing.com/resources/publish-or-perish>
- Scholarometer, from School of Informatics and Computing, Indiana University-Bloomington (currently doesn't work): <http://scholarometer.indiana.edu>
- Scholar Plot (generates plots to visualize an authors academic career, using GS data): <http://scholarplot.com/help.html>
- R Package: scholar: Analyse Citation Data from Google Scholar: <https://cran.r-project.org/web/packages/scholar/index.html>
- R Package: scholarnetwork: Extract and Visualize Google Scholar Collaboration Networks: <https://github.com/pablobarbera/scholarnetwork>
- R Package: cv (builds a list of publications by an author by extracting data from GS): <https://github.com/bomeara/cv>
- R Package: Web::Scraper::Citations (scrapes data from GS author profiles): <https://github.com/JJ/net-citations-scraper>
- Tutorial: *Put Google Scholar citations on your personal website with R, scholar, ggplot2 and cron*: <https://www.r-bloggers.com/put-google-scholar-citations-on-your-personal-website-with-r-scholar-ggplot2-and-cron/>
- Tutorial: *Google scholar scraping with rvest package*: <https://datascienceplus.com/google-scholar-scraping-with-rvest/>
- Tutorial: *Scraping Google Scholar to write your PhD literature chapter*: <https://mystudentvoices.com/scraping-google-scholar-to-write-your-phd-literature-chapter-2ea35f8f4fa1>



MY WORKFLOW (1)

STRUCTURE OF A GOOGLE SCHOLAR RECORD

- No magic recipe. I have to deal with CAPTCHAs and scrape the raw HTML just like everyone else.
- Query embedded in URL: https://scholar.google.com/scholar_lookup?doi=10.1002/asi.23056

The Google Scholar experiment: How to index false papers and manipulate bibliometric indicators

[PDF] arxiv.org

E Delgado López-Cózar... - Journal of the ..., 2014 - Wiley Online Library

Google Scholar has been well received by the research community. Its promises of free, universal, and easy access to scientific literature coupled with the perception that it covers the social sciences and the humanities better than other traditional multidisciplinary databases have contributed to the quick expansion of Google Scholar Citations and Google Scholar Metrics: 2 new bibliometric products that offer citation data at the individual level and at journal level. In this article, we show the results of an experiment undertaken to analyze ...

☆ ⓘ Cited by 101 Related articles All 15 versions Web of Science: 49

```
Elements Console Sources Network Performance Memory Application Security Audits Adblock Plus
▼ <h3 class="gs_rt" ontouchstart="gs_evt_dsp(event)">
  ▼ <a href="http://onlinelibrary.wiley.com/doi/10.1002/asi.23056/full" data-clk="hl=en&sa=T&ct=res&cd=0&ei=WcCFW9rQLcK0mAHi8I7QDA">
    "The Google Scholar experiment: How to index false papers and manipulate bibliometric indicators"
  </a>
</h3>
▼ <div class="gs_a">
  <a href="/citations?user=kyTH0h0AAAAJ&hl=en&oi=sra">E Delgado López-Cózar</a>
  "...&nbsp;- Journal of the&nbsp;..., 2014 - Wiley Online Library"
</div>
▶ <div class="gs_rs">...</div>
▼ <div class="gs_fl">
  ▶ <a href="javascript:void(0)" class="gs_or_sav" title="Save" role="button">...</a>
  ▶ <a href="javascript:void(0)" class="gs_or_cit gs_nph" title="Cite" role="button" aria-controls="gs_cit" aria-haspopup="true">...</a>
  <a href="/scholar?cites=11585142314099351040&as_sdt=2005&scioldt=0,5&hl=en">Cited by 101</a>
  <a href="/scholar?q=related:AB490I2xxqAJ:scholar.google.com/&hl=en&as_sdt=0,5">Related articles</a>
  <a href="/scholar?cluster=11585142314099351040&hl=en&as_sdt=0,5" class="gs_nph">All 15 versions</a>
  <a href="http://gateway.webofknowledge.com/gateway/Gateway.cgi?GWVersion=2&SrcA...L=https://scholar.google.com/&SrcDesc=Back+to+Google+Sc
  "hl=en&sa=T&ct=wosc&cd=0&ei=WcCFW9rQLcK0mAHi8I7QDA" class="gs_nta gs_nph">Web of Science: 49</a> == $0
  ▶ <a href="javascript:void(0)" title="More" class="gs_or_mor gs_ota gs_oph" role="button">...</a>
  ▶ <a href="javascript:void(0)" title="Fewer" class="gs_or_nvi gs_or_mor" role="button">...</a>
</div>
</div>
```



MY WORKFLOW (2)

DATA EXTRACTION PROCESS

- Two-step process (+ cleaning):
 1. Getting raw HTML:
 - [Python](#) script that reads list of queries
 - [Selenium](#) Webdriver (a headless browser doesn't work because it is necessary to solve CAPTCHAs)
 - Pagination is taken into account (Google Scholar displays a maximum of 10 records per page, and a maximum of 1000 records per query)
 - When a CAPTCHA appears, the script pauses. A human has to solve the CAPTCHA, then the script can resume.
 - Each page is saved as an HTML file.
 - More computers: faster.
 2. Parsing HTML:
 - Once all raw files have been downloaded.
 - Another other Python script, using [Scrapy](#) library, reads these files.
 - Using [Xpath](#), relevant data is identified within the HTML.
 - Data is saved to csv file.
 3. Cleaning: getting more complete metadata from source website and/or CrossRef...



DATA PRODUCTS (1)

BUILDING NEW TOOLS ON TOP OF GOOGLE SCHOLAR DATA

The screenshot shows the 'JOURNAL SCHOLAR METRICS' website. The header includes the title 'JOURNAL SCHOLAR METRICS' and the subtitle 'ARTS, HUMANITIES, AND SOCIAL SCIENCES'. There are navigation links for HOME, ABOUT, METHODOLOGY, OUR TEAM, OTHER PROJECTS, and FAQ. A search bar is present with the placeholder text 'Search a journal'. The main content area is divided into two sections: 'SUBJECT CATEGORY RANKINGS' and 'COUNTRY RANKINGS'. The 'SUBJECT CATEGORY RANKINGS' section is currently displaying 'SOCIAL SCIENCES' and lists various sub-fields with their respective journal counts. The 'COUNTRY RANKINGS' section shows a world map with a color scale from light blue to dark blue, indicating the number of journals from each country. The footer contains the text: 'Journal Scholar Metrics is a product developed by EC3 Research Group: Evaluación de la Ciencia y la Comunicación Científica. Universidad de Granada. Campus de Cartuja s/n. Granada (Spain).'

SUBJECT CATEGORY RANKINGS	
SOCIAL SCIENCES	
ANTHROPOLOGY	(298)
COMMUNICATION	(320)
BUSINESS, ECONOMICS & MANAGEMENT	(1761)
EDUCATION	(1126)
GEOGRAPHY & URBAN STUDIES	(548)
LAW	(920)
LIBRARY & INFORMATION SCIENCE	(277)
POLITICAL SCIENCE, ADMINISTRATION & INTERNATIONAL RELATIONS	(1074)
PSYCHOLOGY	(1032)
SOCIOLOGY	(1007)
MULTIDISCIPLINARY	(202)
SOCIAL WORK	(132)
SPORT SCIENCES	(213)

ARTS & HUMANITIES

JOURNAL SCHOLAR METRICS

Presents bibliometric indicators for 9,196 journals in the Arts, Humanities, and Social Sciences, by discipline.

These areas have traditionally presented more difficulties in terms of bibliometric assessment, mainly because of the lack of international, geographically and linguistically unbiased tools


<http://www.journal-scholar-metrics.infoec3.es>



UNIVERSIDAD
DE GRANADA


DATA PRODUCTS (1)

BUILDING NEW TOOLS ON TOP OF GOOGLE SCHOLAR DATA




Scholar Mirrors


Bibliometrics, Scientometrics, Informetrics, Webometrics, and Altmetrics
in Google Scholar Citations, ResearcherID, Researchgate, Mendeley, and Twitter




HOME
ABOUT
METHODOLOGY
OUR TEAM
OTHER PROJECTS




AUTHORS



DOCUMENTS



JOURNALS



PUBLISHERS

General overview

Displaying core authors 1-20 of 398. Sorted by GS citations (last 5 years), decreasingly. Check to display related authors as well

Search an author

Name	Online presence	Google Scholar		ResearcherID		ResearchGate		Mendeley		Twitter	
		Citations	H Index	Citations	H Index	RG Score	Downloads	Readers	Followers	Tweets	Followers
Loet Leydesdorff		26484	73	6444	44	45.14	32165	0	11	84	375
Eugene Garfield*		22622	55	8790	153	-	-	-	-	-	-
Mike Thelwall		13840	61	3593	32	42.64	24989	7423	36	85	522
Derek J. de Solla Price		13263	33	-	-	-	-	-	-	-	-
Francis Narin		11297	45	-	-	32.38	795	-	-	-	-
Wolfgang Glänzel		10796	54	4924	38	41.16	10572	-	-	-	-
Ronald Rousseau		9570	42	NA	NA	42.75	8066	-	-	-	-
Chaomei Chen		9512	43	1740	20	34.65	31579	985	3	67	65
Anthony (Ton) F.J. van Raan		9200	53	-	-	38.47	6014	-	-	58	168
Ben R. Martin		8975	39	-	-	-	-	-	-	-	-
András Schubert		8655	45	4121	31	39.24	1962	-	-	-	-
Peter Ingwersen		8356	35	NA	NA	30.64	8600	-	-	-	-
Henk F. Moed		8256	46	-	-	-	-	-	-	-	-
Blaise Cronin		7347	43	-	-	33.9	1891	-	-	-	-
Henry Small		7307	32	3360	23	-	-	-	-	-	-
Tibor Braun		7231	41	NA	NA	NA	NA	-	-	-	-
Vasily V. Nalimov		6343	31	-	-	-	-	-	-	-	-
Lutz Bornmann		6108	40	2676	27	43.12	13556	0	0	405	240
Belver C. Griffith		5695	26	-	-	-	-	-	-	-	-
Howard D. White		5569	30	NA	NA	29.58	3376	0	0	-	-

First | Previous | Next | Last

SCHOLAR MIRRORS

Bibliometric and altmetric indicators for authors, documents, journals, and publishers in the field of Bibliometrics, Scientometrics, Informetrics, Webometrics, and Altmetrics in Google Scholar Citations, ResearcherID, Researchgate, Mendeley, and Twitter.

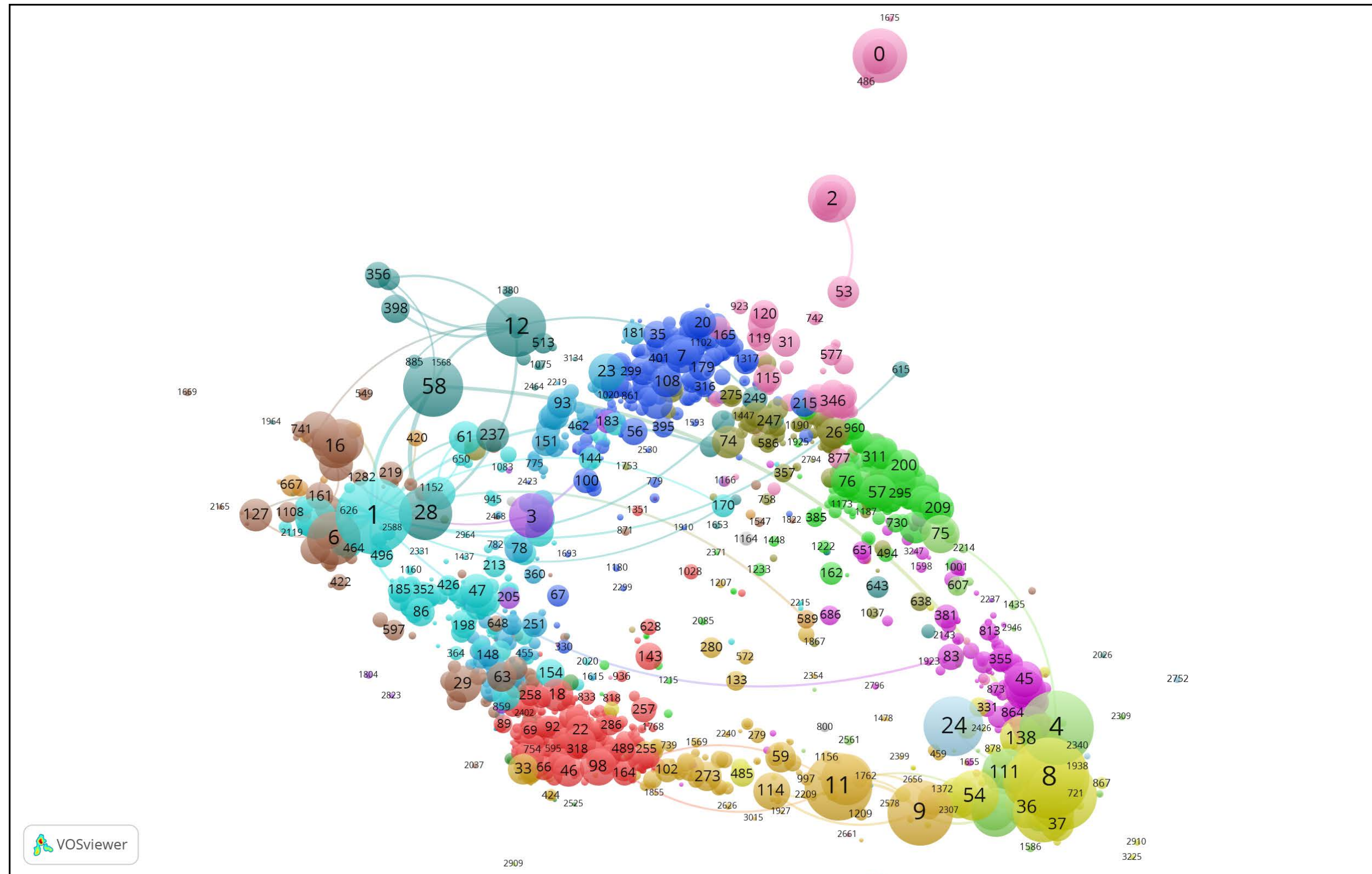
Data was extracted from Google Scholar, ResearcherID, ResearchGate, Mendeley, and Twitter.

<http://www.scholar-mirrors.infoec3.es>



DATA PRODUCTS (1)

BUILDING NEW TOOLS ON TOP OF GOOGLE SCHOLAR DATA



WORK IN PROGRESS

Google Scholar-powered scientific information system that displays data about all researchers working in Spain.

Dataset: 44,500 profiles in Google Scholar Citations (profile service) → 2 million documents → approx. 30 million citations

Necessary: generating a document-level classification for this collection of GS data. I'm using Ludo Waltman's and Nees Jan van Eck's smart local moving algorithm.



UNIVERSIDAD
DE GRANADA

ONE LAST THOUGHT

- If you make data available, people will build on top of it



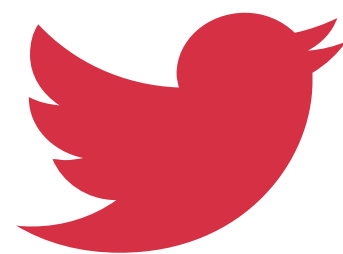
- We want to provide a glimpse of what could be possible to do with Google Scholar data, if it stopped being a black box

THANK YOU FOR YOUR ATTENTION

✉ albertomartin@ugr.es

✉ edelgado@ugr.es

NOT JUST
Google
scholar's
Digest



@[GScholarDigest](https://twitter.com/GScholarDigest)