UNIVERSIDAD DE GRANADA

Focus On Open Science

Google Scholar: Highly Comprehensive Coverage… Inside a Hermetic Black Box

Alberto Martín-Martín

Budapest, September 17th, 2019

THE NEED OF OPEN RESEARCH METADATA

“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

MENU

nature
International journal of science

EDITORIAL · 11 SEPTEMBER 2019

Set citation data free

https://www.nature.com/articles/d41586-019-02669-3

THE ROAD FROM CLOSED TO OPEN RESEARCH METADATA

Clarivate Analytics
WEB OF SCIENCE™

Only one until 2004
Closed

ELSEVIER
Scopus

2004-
Closed

Google Scholar

2004-
Free, not open (no API)

Microsoft Academic

2016-
Free, freemium API
Open dump of db

Dimensions

2018-
Freemium, API
(Free for research)

OpenCitations

2010-
Completely Open
(CC-0)

NO DATABASE IS COMPLETE

Five Blind Men and an Elephant

- John Godfrey Saxe

Mike Kline. CC-BY 2.0

Thanks to Dr. Elizabeth Gadd for discovering this poem to me in one of her insightful posts about responsible metrics

INTRODUCTION TO GOOGLE SCHOLAR

2004: GOOGLE SCHOLAR LAUNCH

- Free
- **Inclusive (vs. selective)** indexing
- Citation data
- Access to full text (if available)
- GOAL: facilitate content discovery

SINCE 2005: WIDELY USED

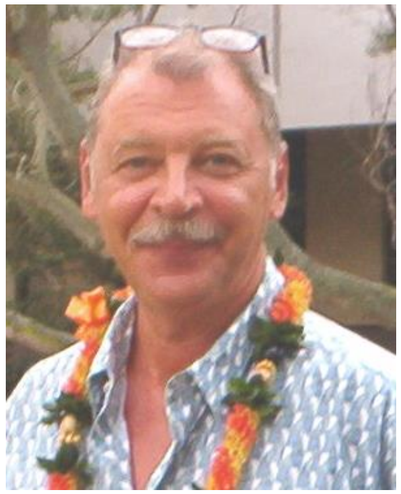
- Main source of traffic to journals
- Preferred starting point for literature search

2007: LAUNCH OF HARZING'S *PUBLISH OR PERISH*

- Facilitates citation analysis (no longer limited to people with access to WoS/Scopus)

SINCE 2005: CRITICISM

- Coverage gaps
- Unreliable citation counts
- Errors in bibliographic data



SINCE 2007: CONSOLIDATION

- More publishers join
- Studies report broader coverage
  Many bibliographic errors are fixed

2011, 2012: SPIN-OFF SERVICES

- GS Citations (author profiles)
- GS Metrics (journal rankings)

2014: TENTH ANNIVERSARY

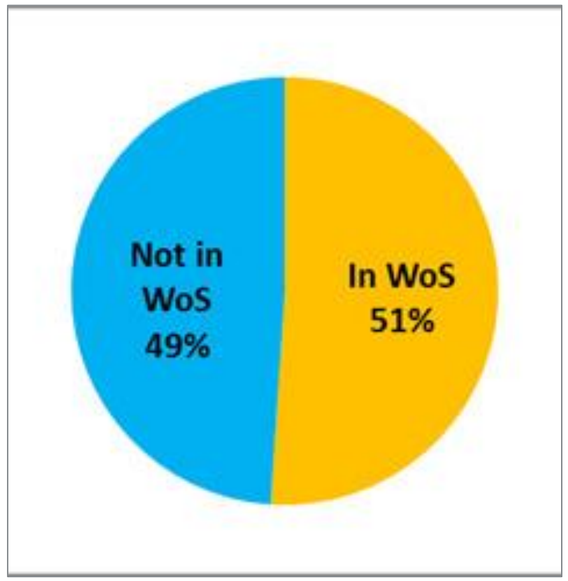
- Citation counts easy to game
- Size: 114-160 million documents
- My doctoral training starts…

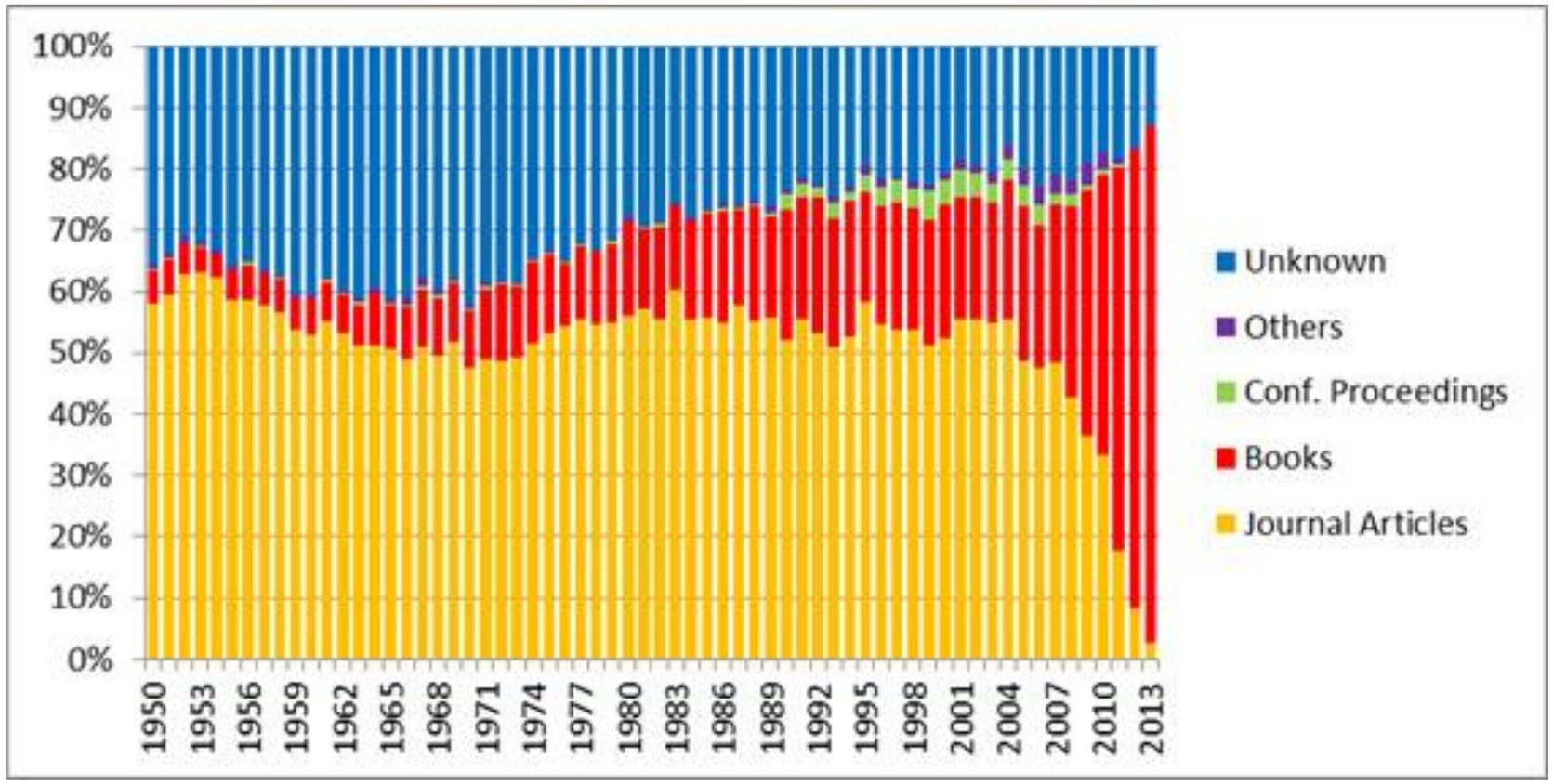
GOOGLE SCHOLAR: AN ATYPICAL GOOGLE PRODUCT

- Not commercially exploited
  - It doesn't display ads
  - "Google Scholar does not currently make money" (interview to GS's chief engineer Anuarg Acharya)
- Are we paying with our data?
  - Anurag Acharya: [we] "don't actually track past searches by specific researchers"
  - Unlike in other Google products, no privacy notice when accessing GS (compulsory in Europe with GDPR).
  - GS not present in Google's dashboard of information collected about an user.

GOOGLE SCHOLAR AS A SOURCE OF DATA

First exploratory analysis:
Analysis of 64,000 documents published in 1950-2013

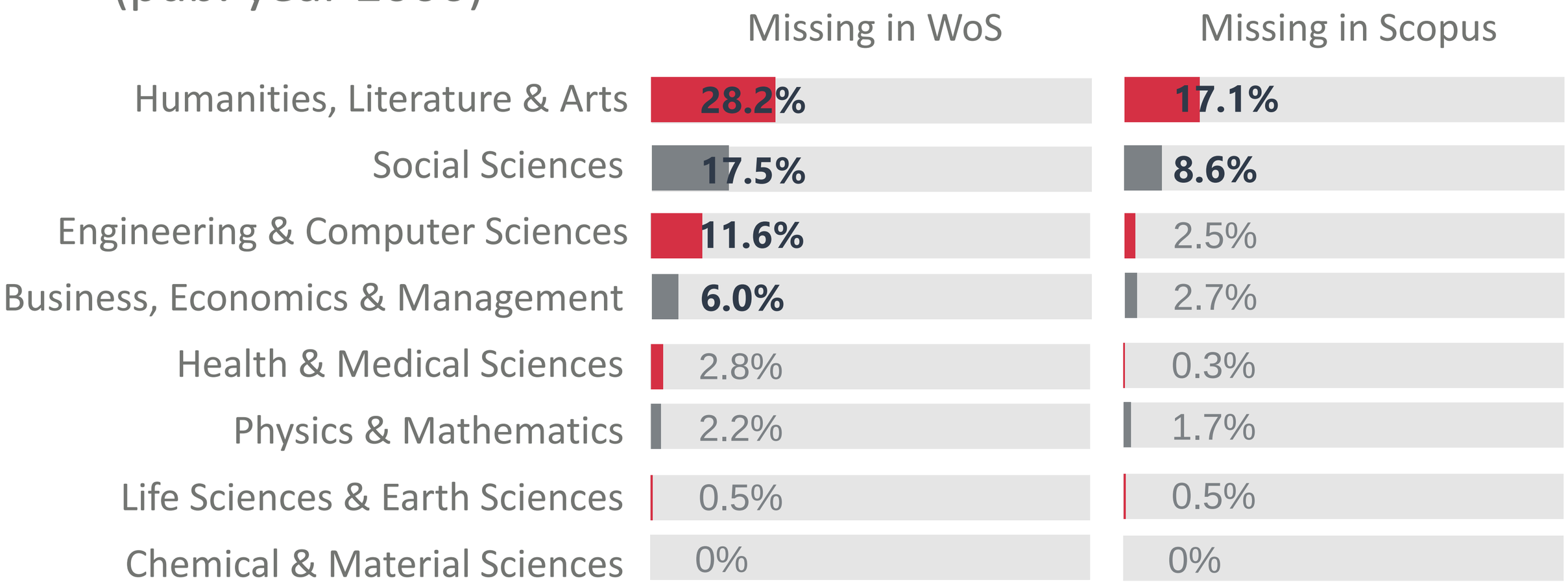




Martín-Martín, A., Orduna-Malea, E., Ayllón, J. M., & Delgado López-Cózar, E. (2016). A two-sided academic landscape: snapshot of highly-cited documents in Google Scholar (1950-2013). *Revista Española de Documentacion Cientifica*, *39*(4), e149. https://doi.org/10.3989/redc.2016.4.1405

GOOGLE SCHOLAR AS A SOURCE OF DATA

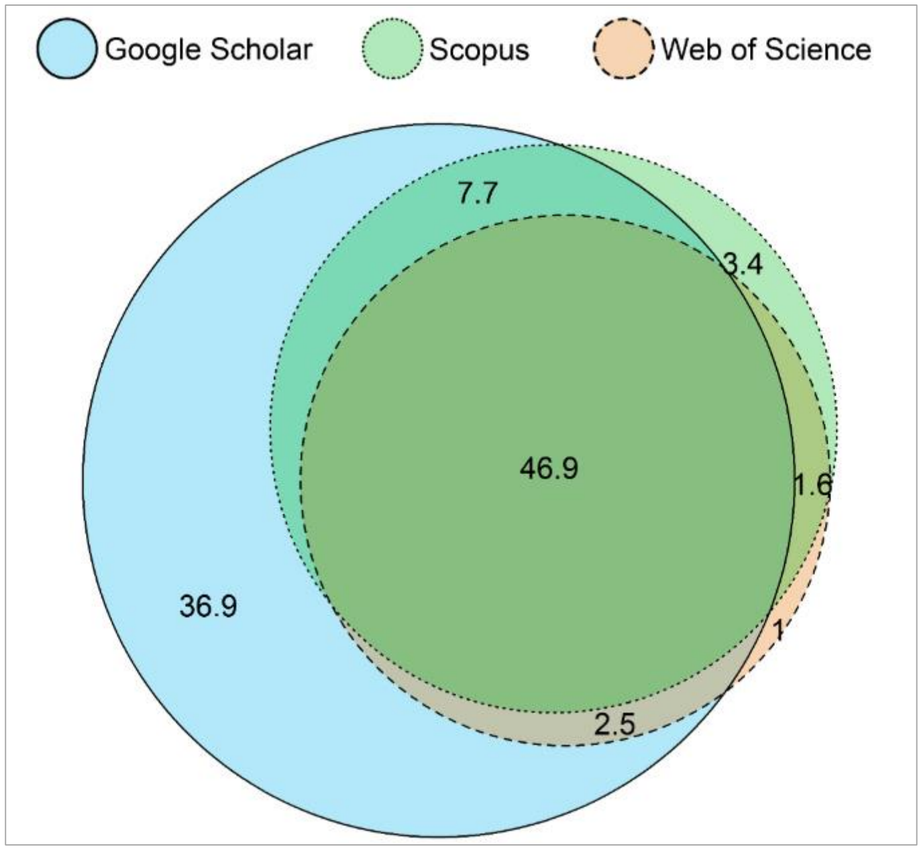
Analysis of highly-cited documents:
Top 10 most cited documents in GS, across 252 subject categories (pub. year 2006)

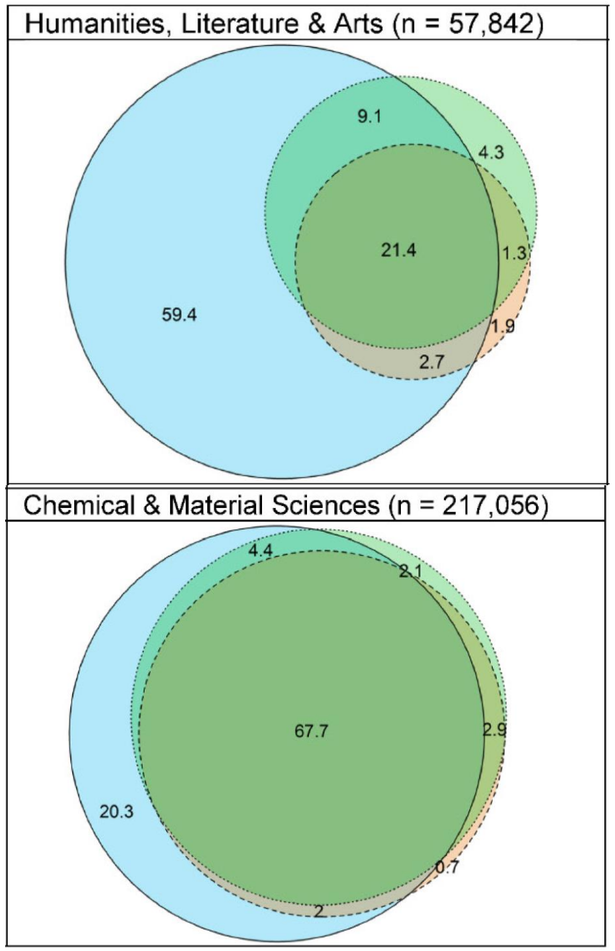


Martín-Martín, A., Orduna-Malea, E., & Delgado López-Cózar, E. (2018). Coverage of highly-cited documents in Google Scholar, Web of Science, and Scopus: a multidisciplinary comparison. *Scientometrics*, *116*(3), 2175–2188. https://doi.org/10.1007/s11192-018-2820-9

GOOGLE SCHOLAR AS A SOURCE OF DATA

Analysis of citations:
2,448,055 citations to 2,299 highly-cited articles across 252 subject categories

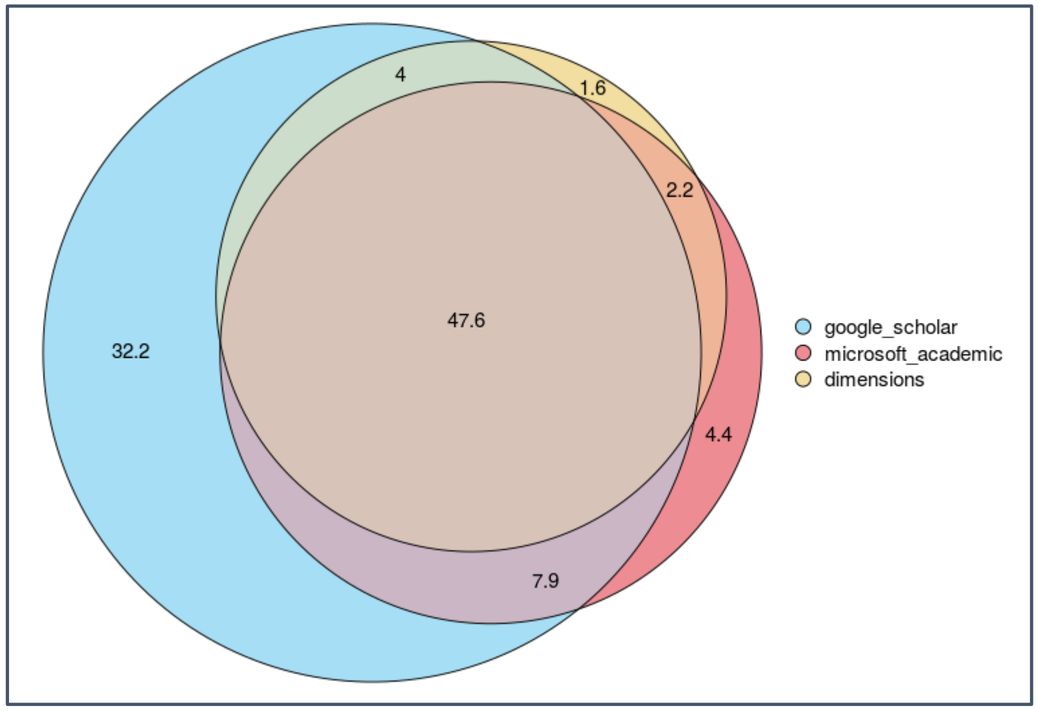


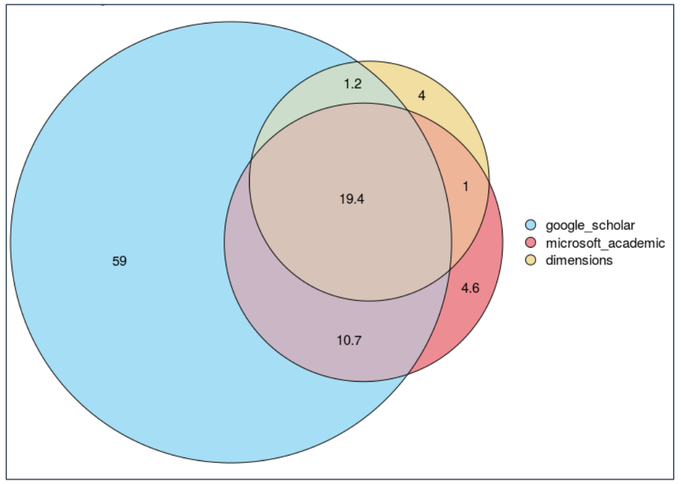


Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & Delgado López-Cózar, E. (2018). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, *12*(4), 1160–1177. https://doi.org/10.1016/J.JOI.2018.09.002

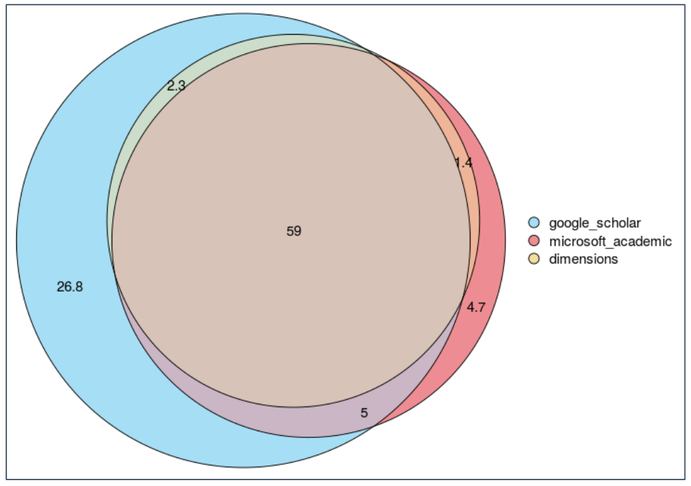
GOOGLE SCHOLAR AS A SOURCE OF DATA

New study: same sample, updated data to June 2019 and added three more sources for comparison: Microsoft Academic, Dimensions, and COCI (CrossRef data)





History

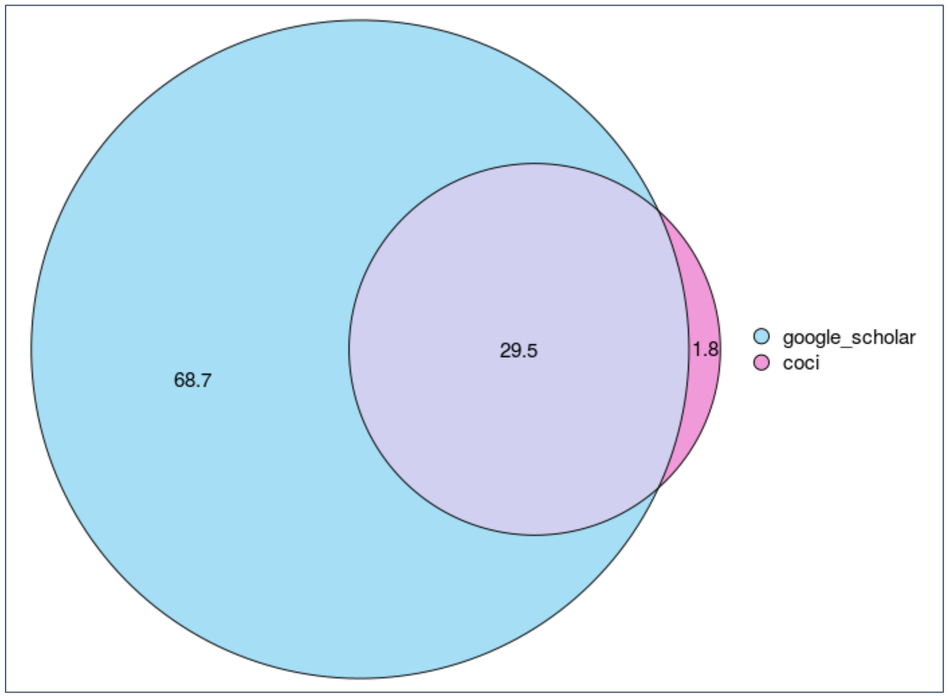


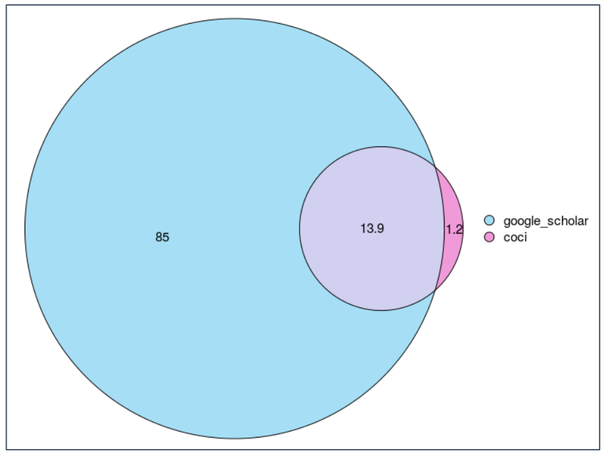
Molecular Biology

Exclusive!
Preliminary results

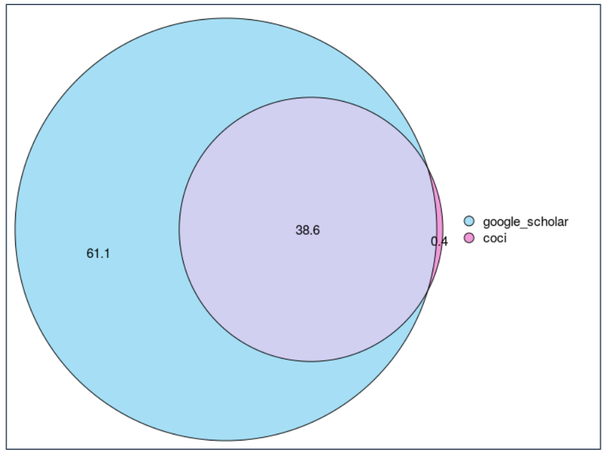
GOOGLE SCHOLAR AS A SOURCE OF DATA

New study: same sample, updated data to June 2019 and added three more sources for comparison: Microsoft Academic, Dimensions, and COCI (CrossRef data)





Chinese Studies & History

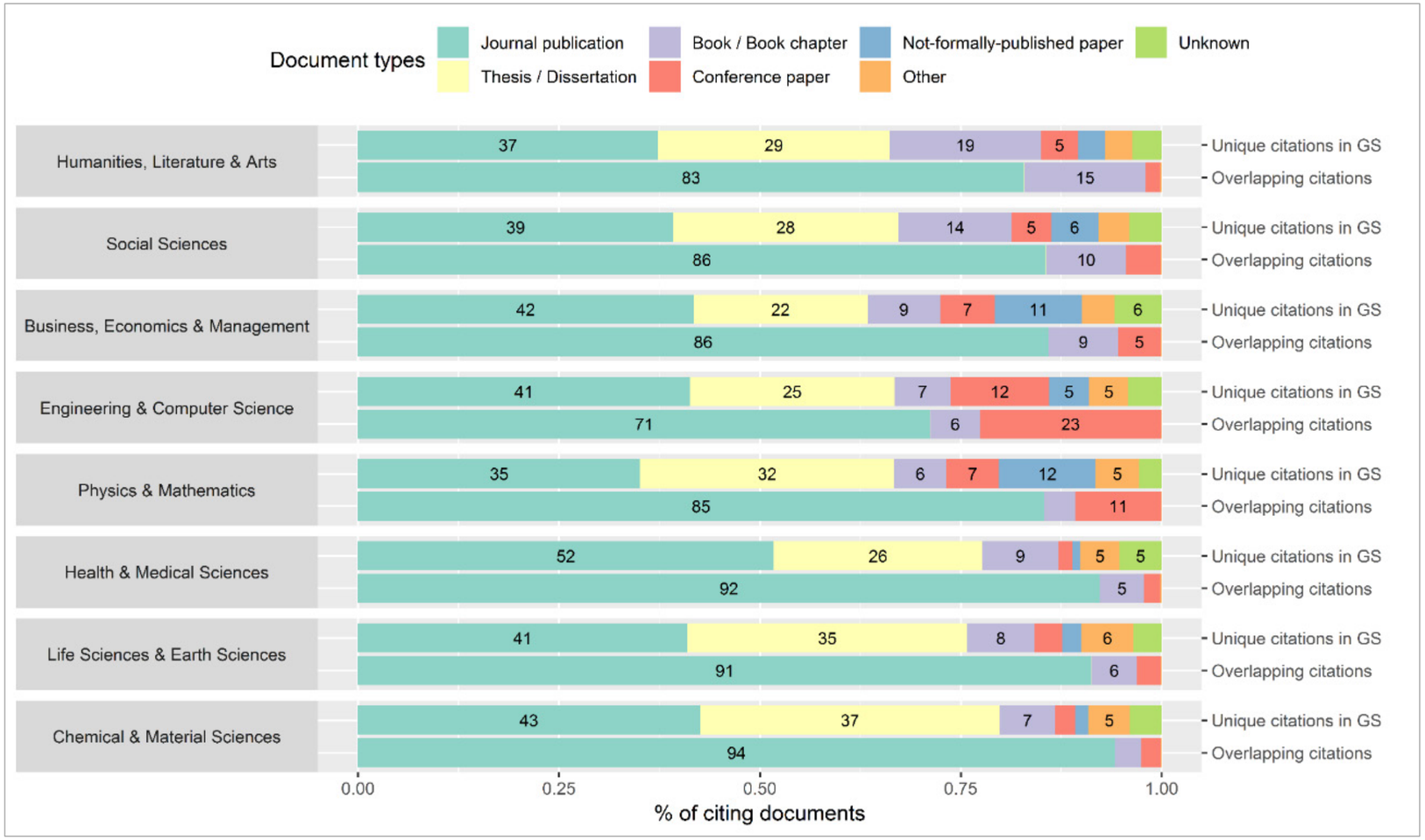


Crystallography & Structural Chemistry

Exclusive!
Preliminary results

GOOGLE SCHOLAR AS A SOURCE OF DATA

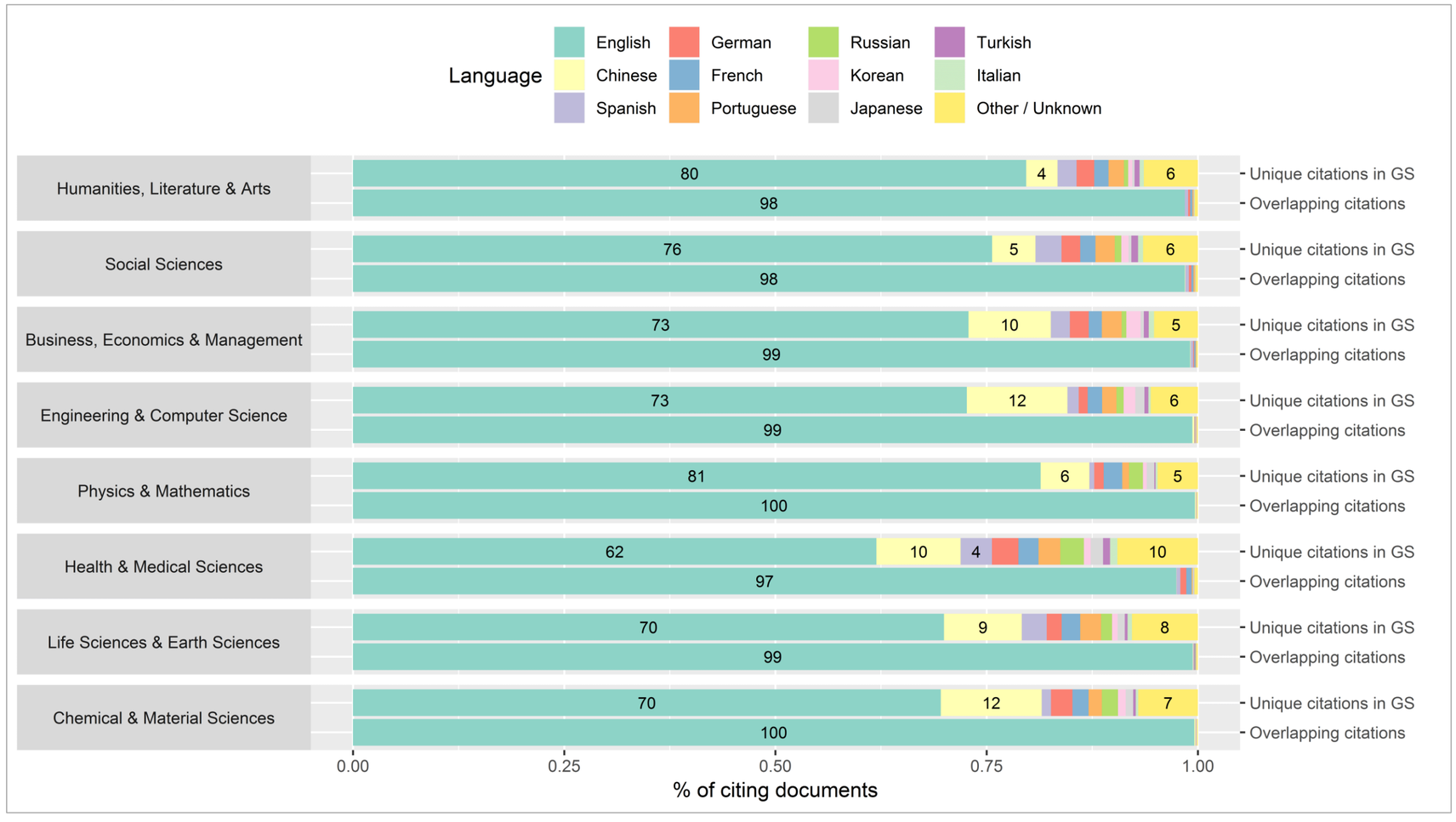
Analysis of citations:
2,448,055 citations to 2,299 highly-cited articles across 252 subject categories



Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & Delgado López-Cózar, E. (2018). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, *12*(4), 1160–1177. https://doi.org/10.1016/J.JOI.2018.09.002

GOOGLE SCHOLAR AS A SOURCE OF DATA

Analysis of citations:
2,448,055 citations to 2,299 highly-cited articles across 252 subject categories



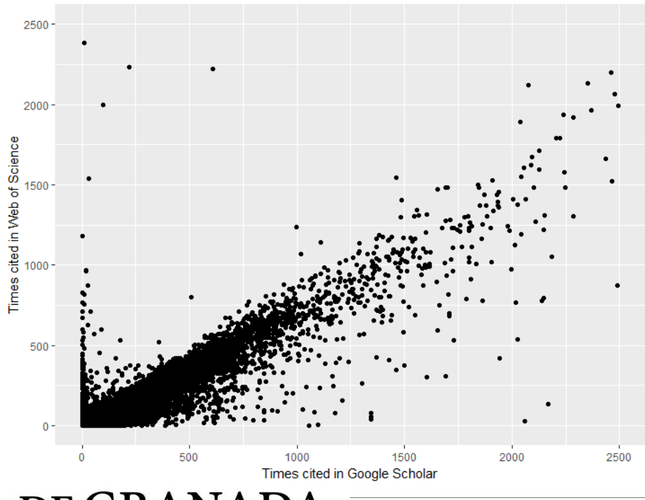
Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & Delgado López-Cózar, E. (2018). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, *12*(4), 1160–1177. https://doi.org/10.1016/J.JOI.2018.09.002

GOOGLE SCHOLAR AS A SOURCE OF DATA

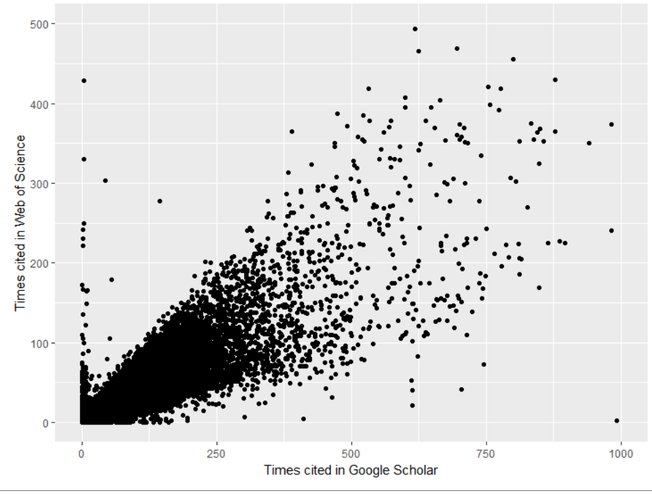
Correlations of citation counts

Document-level citation counts				
Date of data collection	GS-WoS N docs	GS-WoS Spearman correlation*	GS-Scopus N docs	GS-Scopus Spearman correlation
April-May 2018	1.03 million	**0.94 (0.78-0.98)**	1.2 million	**0.96 (0.93-0.99)**
February 2017	69,261	**0.91**		
June-October 2016	2.26 million	**0.91**		
July 2015	1,055	**0.76**		
July 2015	150	**0.80**		
February 2015	239	**0.63**		
May 2014	32,679	**0.73**		

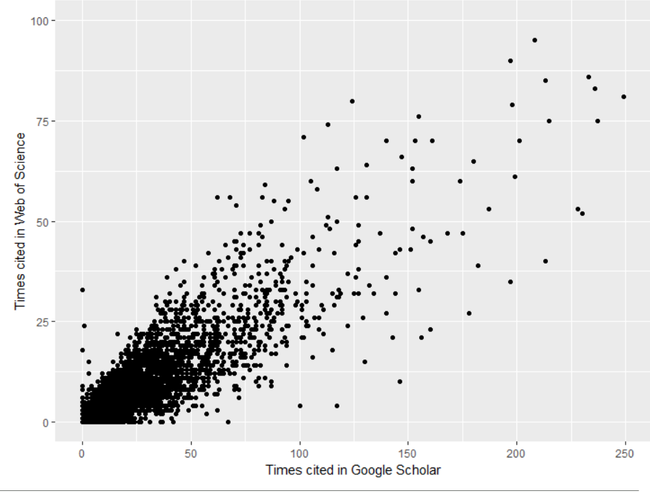
Sciences



Social Sciences



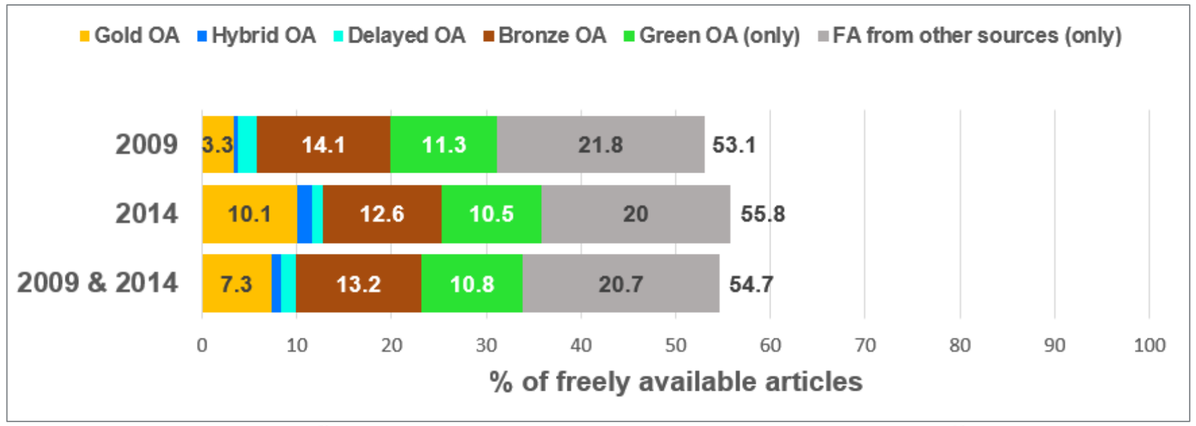
Arts & Humanities

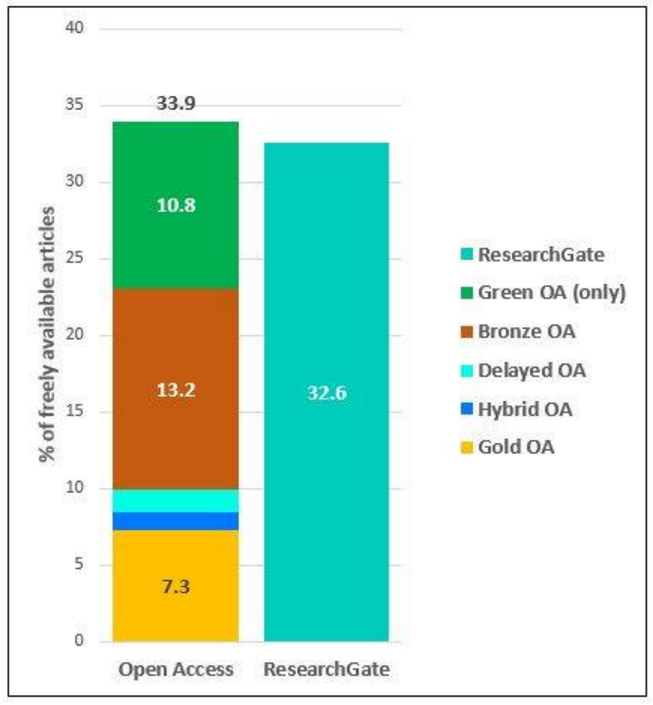


DE GRANADA

GOOGLE SCHOLAR AS A SOURCE OF DATA

Open Access data:
2.26 million WoS-sourced documents were searched in GS





Country	Documents	% OA from publisher	% OA from repositories*	% OA Total	% FA other sources†	% OA + FA†
World	1,331,795	25.3	10.5	35.8	20.0	55.7
USA	360,889	29.1	18.2	47.3	18.9	66.2
Peoples R China	231,162	22.9	4.3	27.2	18.7	46.0
Germany	96,265	28.6	13.4	42.0	19.2	61.3
England	89,996	35.0	15.9	50.9	17.3	68.3
Japan	71,587	26.6	9.9	36.5	13.4	49.9
France	66,648	26.5	17.4	43.9	23.5	67.4
Canada	60,342	28.1	10.5	38.6	23.1	61.7
Italy	58,397	26.2	11.9	38.1	25.6	63.7
Australia	53,822	26.2	10.5	36.7	24.9	61.7
Spain	51,586	25.3	13.9	39.2	24.7	63.9
South Korea	51,036	26.2	5.4	31.6	17.9	49.5
India	50,468	15.7	7.4	23.1	25.6	48.7
Netherlands	36,228	33.7	14.2	47.9	22.9	70.8
Brazil	34,517	37.0	8.8	45.8	25.8	71.6
Russia	28,108	10.6	9.7	20.3	23.9	44.3

Martín-Martín, A., Costas, R., van Leeuwen, T., & Delgado López-Cózar, E. (2018). Evidence of open access of scientific publications in Google Scholar: A large-scale analysis. *Journal of Informetrics*, *12*(3), 819–841. https://doi.org/10.1016/j.joi.2018.06.012

GOOGLE SCHOLAR AS A SOURCE OF DATA

Taxonomy of errors in GS:

- Coverage errors
  - False positives/negatives
- Parsing errors: incorrect / incomplete metadata
- Matching errors:
  - Source document matching: duplicate records
  - Citation matching: duplicate citations

Errors in GS Citations (author profiles):

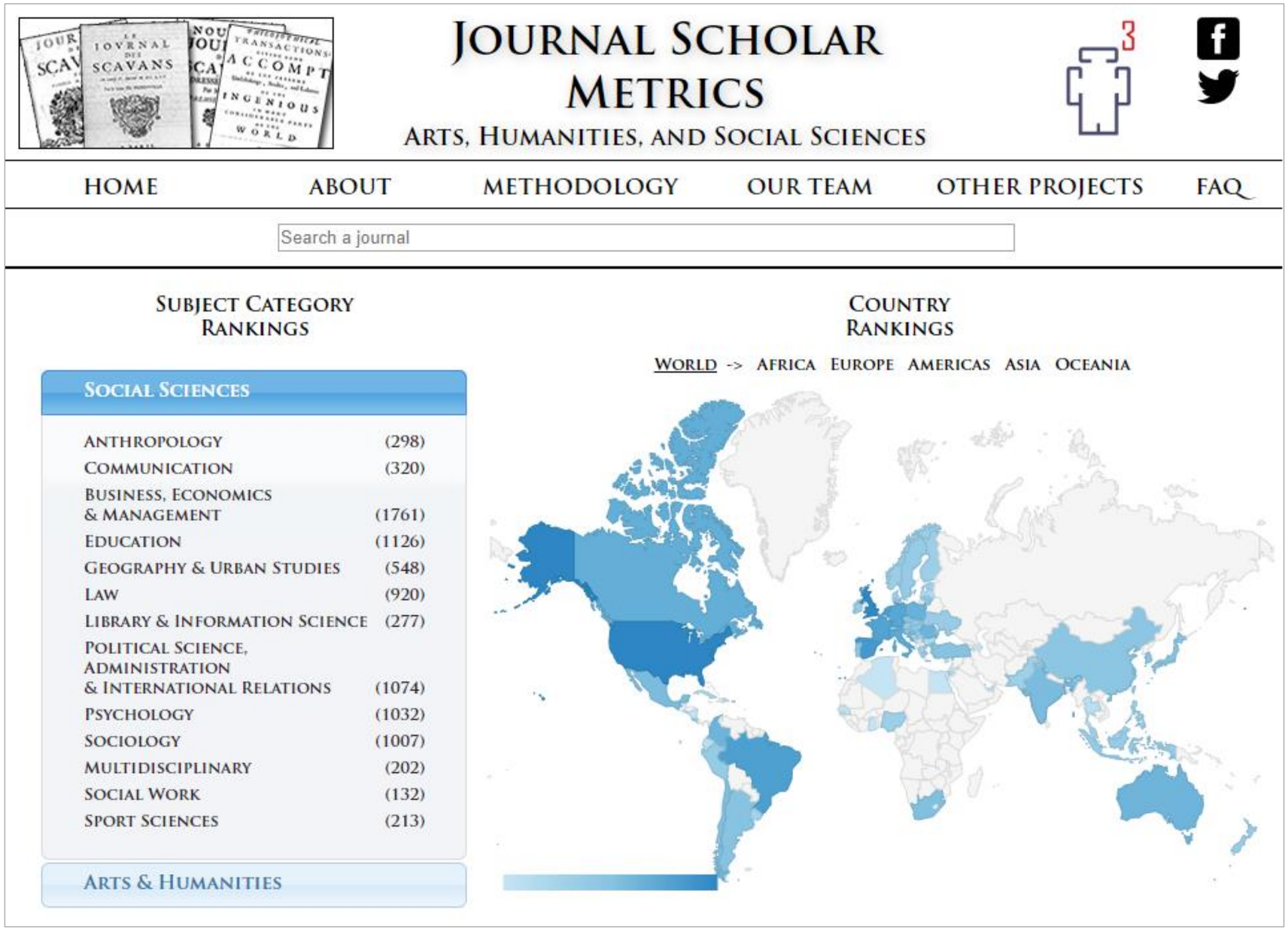
- Duplicate profiles
- Misattributed documents

REUSING DATA FROM GOOGLE SCHOLAR

Journal Scholar Metrics

- 9,196 SSH journals
- Consensus journal classification
- Possible to filter by country of publication

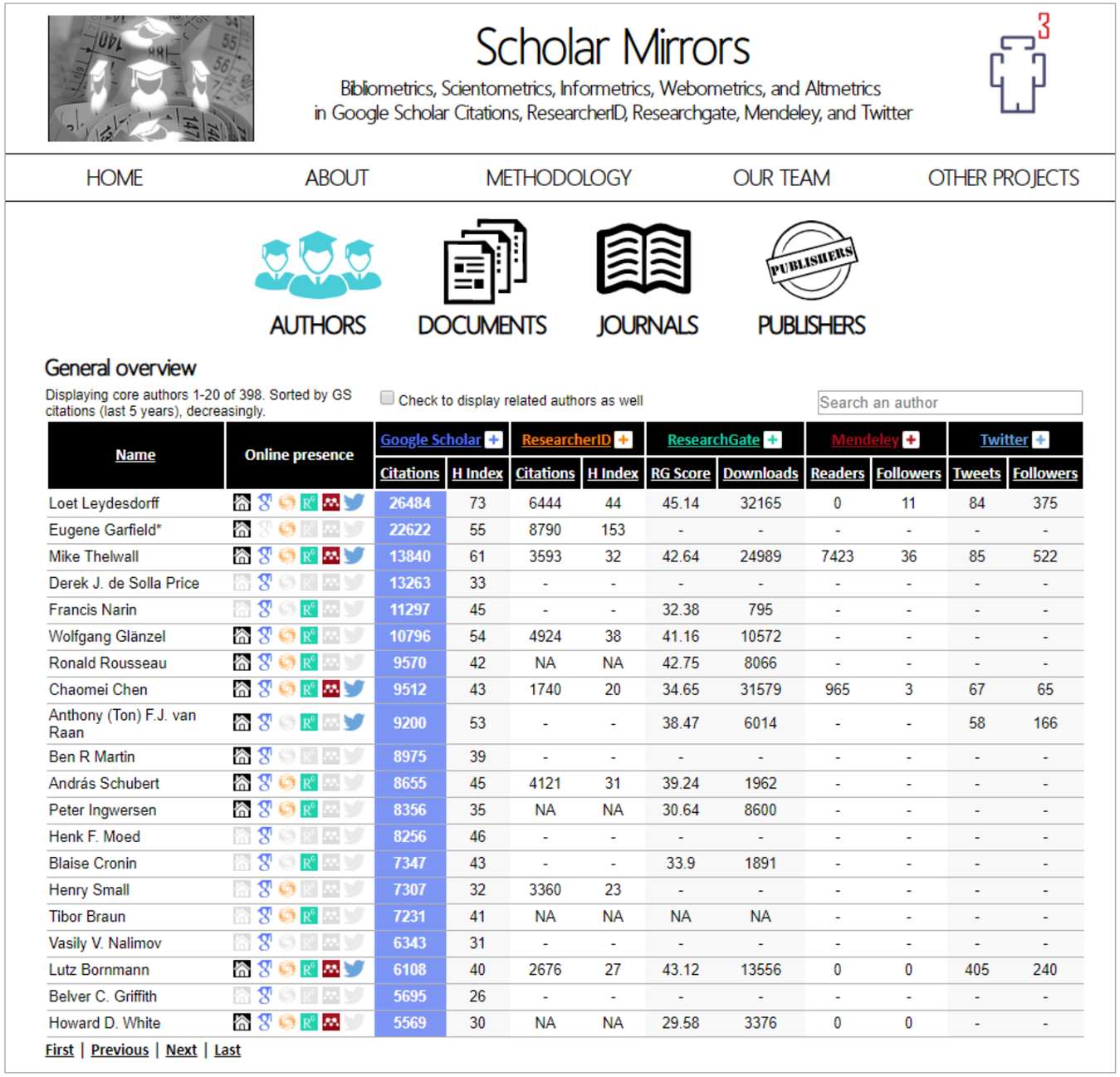
- **Spanish journals: JSM: 861 / 9196 (9%); SJR: 261 / 8180 (3.1%); WoS: 88 / 4166 (2%)**



REUSING DATA FROM GOOGLE SCHOLAR

Scholar Mirrors

- 814 authors
- Multifaceted Analysis (MADAP)
- Different types of indicators from five data sources



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

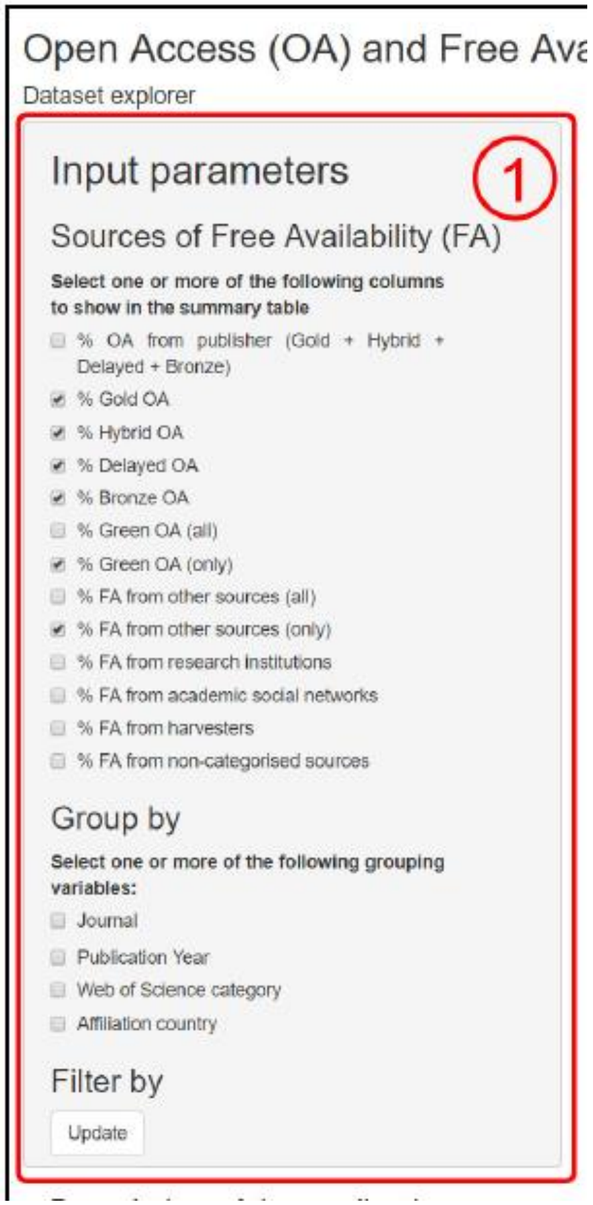
Search an author

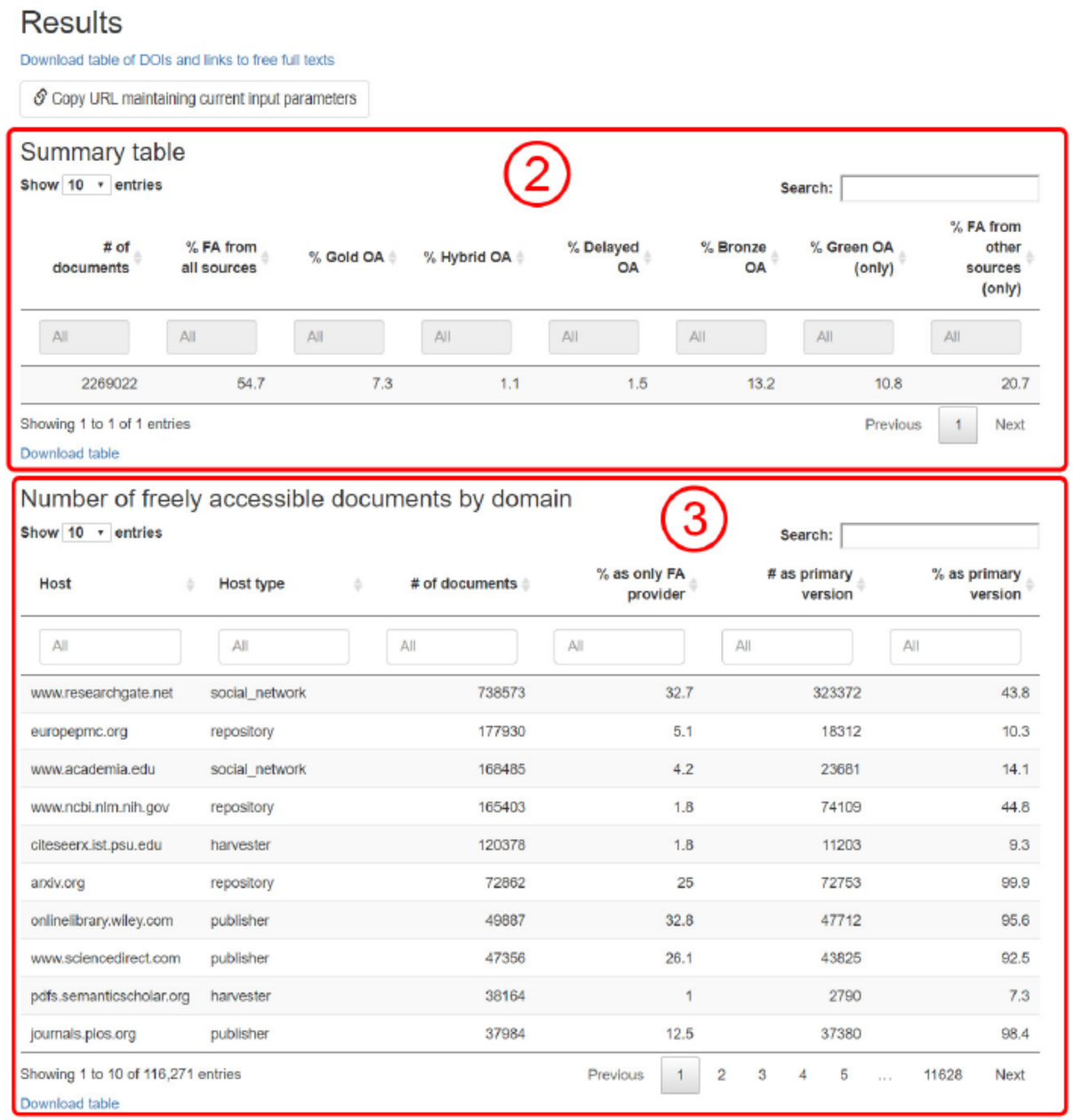
Name	Online presence	Google Scholar		ResearcherID		ResearchGate		Mendeley		Twitter	
		Citations	H Index	Citations	H Index	RG Score	Downloads	Readers	Followers	Tweets	Followers
Loet Leydesdorff		26484	73	6444	44	45.14	32165	0	11	84	375
Eugene Garfield*		22622	55	8790	153	-	-	-	-	-	-
Mike Thelwall		13840	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	965	3	67	65
Anthony (Ton) F.J. van Raan		9200	53	-	-	38.47	6014	-	-	58	166
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

REUSING DATA FROM GOOGLE SCHOLAR

Open Access dashboard





Martín-Martín, Alberto. Creation of bibliometric tools for evaluation based on data from Google Scholar. Granada: Universidad de Granada, 2019. [http://hdl.handle.net/10481/56212]

REUSING DATA FROM GOOGLE SCHOLAR

Enhanced author profiles
(work in progress)

Sample:
>40,000 authors working in Spain
>2 million unique document
>24 million citations

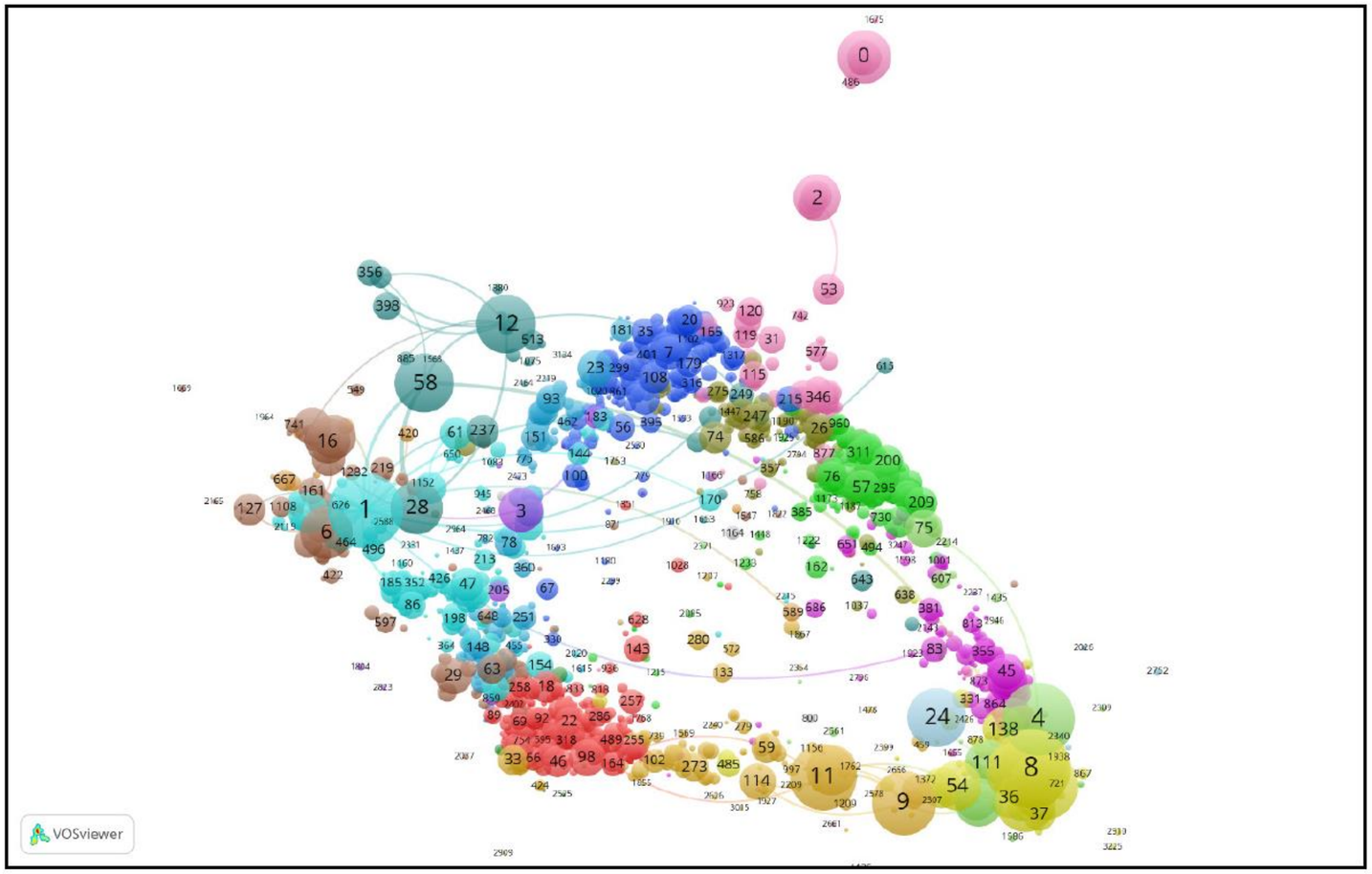


Figure 2. Clusters of documents displayed in the Google Scholar Citations profiles of researchers working in Spain

CONCLUSIONS

STRENGTHS of Google Scholar as a source of data:

- Extensive coverage: almost everything in WoS/Scopus, and more
    - Specially in Arts, Humanitites, and Social Sciences
    - Makes visible document types that have been traditionally excluded from analyses
    - More diverse distribution of languages
- Very high correlations of citation counts, despite unique sources (and errors) in GS
- GS citation data:
    - No significant differences to WoS/Scopus data when analysing STEM fields
    - significantly more useful in SSH.

CONCLUSIONS

LIMITATIONS of Google Scholar as a source of data:

- Lack of **transparency** about size and coverage
- Lack of support for advanced search and filtering
- Dynamic coverage: potential (silent) decrease in coverage
- Limited document metadata
- No options to export data in bulk (necessary to deal with CAPTCHAs manually)
- More open to manipulation than controlled databases

THANK YOU FOR YOUR ATTENTION

albertomartin@ugr.es

@albertomartin