UNIVERSIDAD DE GRANADA

Facultad de
Comunicación y Documentación

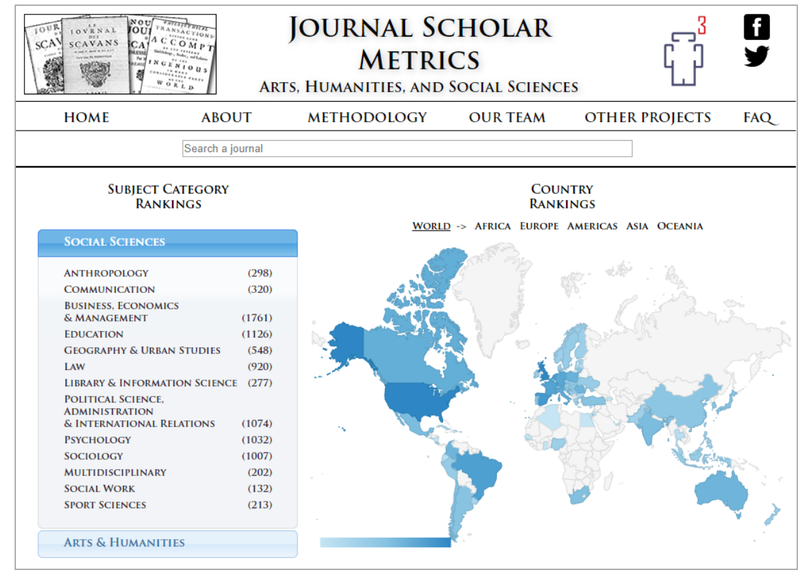
Generación de herramientas de evaluación bibliométrica a partir de Google Scholar

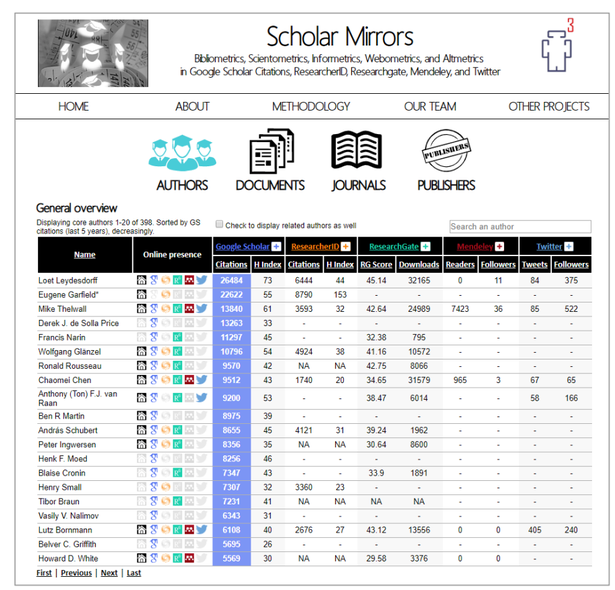
Creation of bibliometric tools for evaluation based on data from Google Scholar

Doctoral thesis defense
Candidate: Alberto Martín-Martín
Advisor: Emilio Delgado López-Cózar

Granada (Spain), June 7th, 2019

THESIS BY COMPILATION OF STUDIES





UNIVERSIDAD DE GRANADA

SUMMARY

1. INTRODUCTION

2. OBJECTIVES

3. GOOGLE SCHOLAR AS A SOURCE OF DATA

4. REUSING DATA FROM GOOGLE SCHOLAR

5. CONCLUSIONS

INTRODUCTION

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

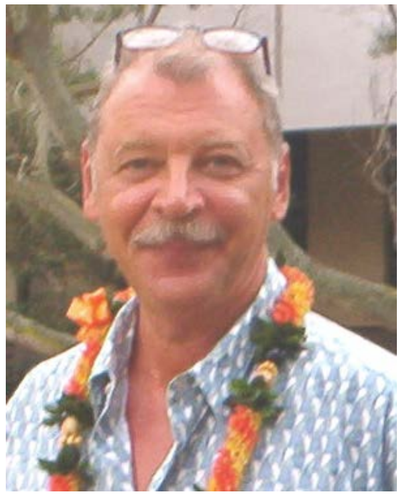
- Main source of traffic to journals
- Preferred starting point for literature search

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…

OBJECTIVES

1

IDENTIFY GENERAL CHARACTERISTICS OF GS AS A SOURCE OF DATA

Coverage
Citation data
Open access data
Errors

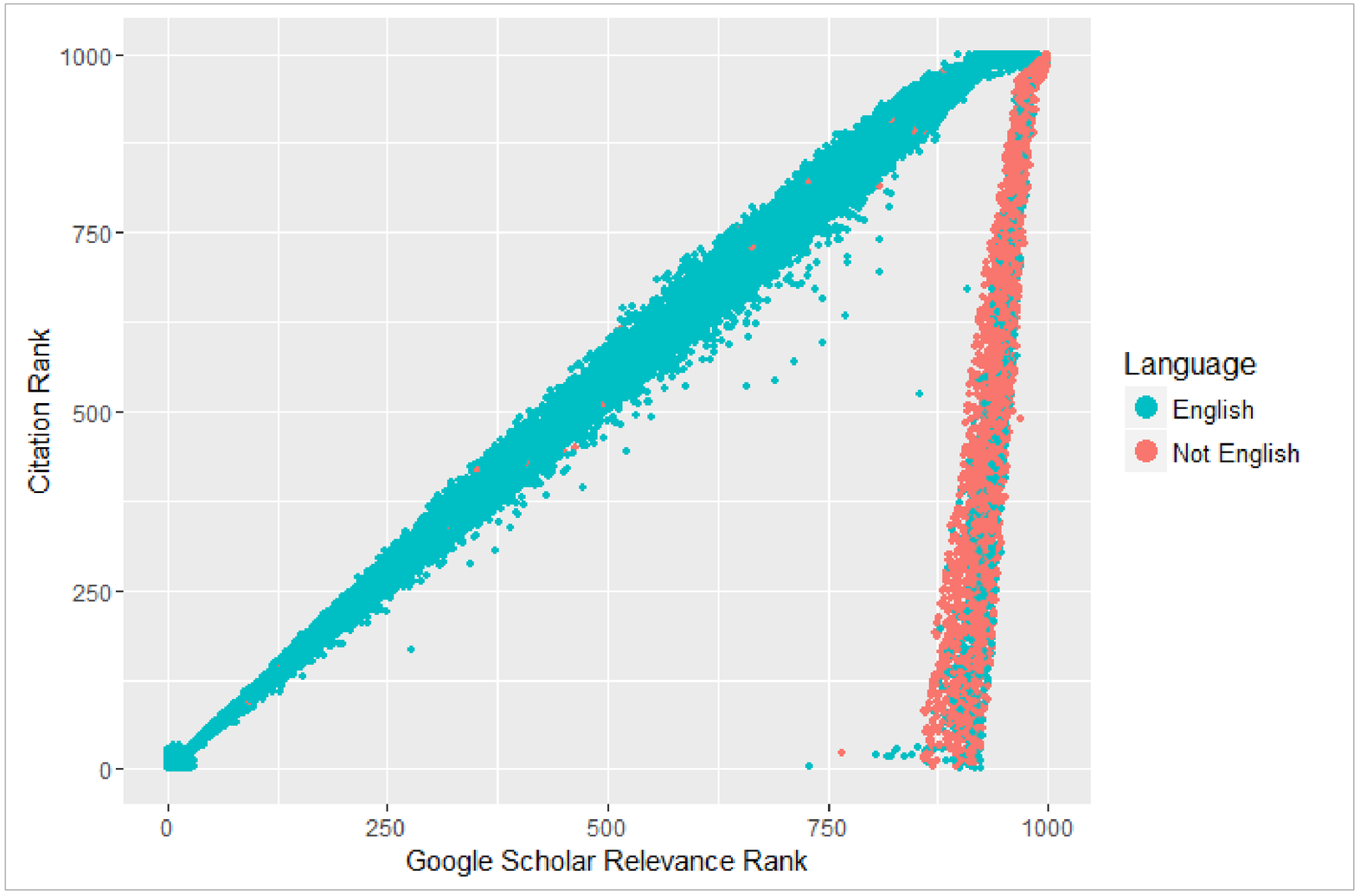
2

DEVELOPMENT OF APPLICATIONS THAT REUSE DATA FROM GS

Journal Scholar Metrics
Scholar Mirrors
Open Access dashboard
Enhanced author profiles

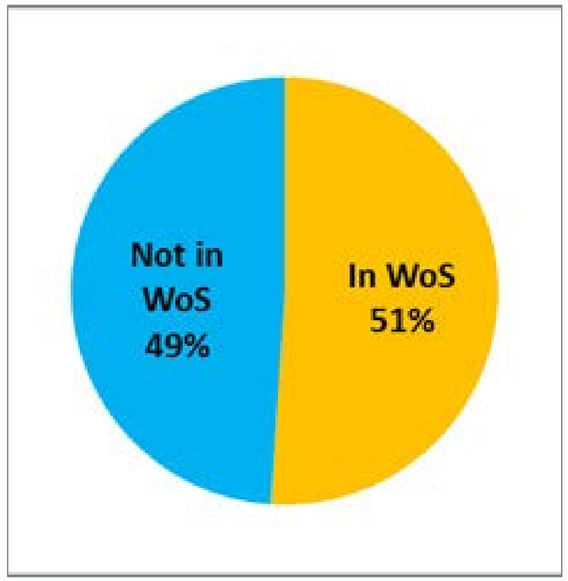
GOOGLE SCHOLAR AS A SOURCE OF DATA

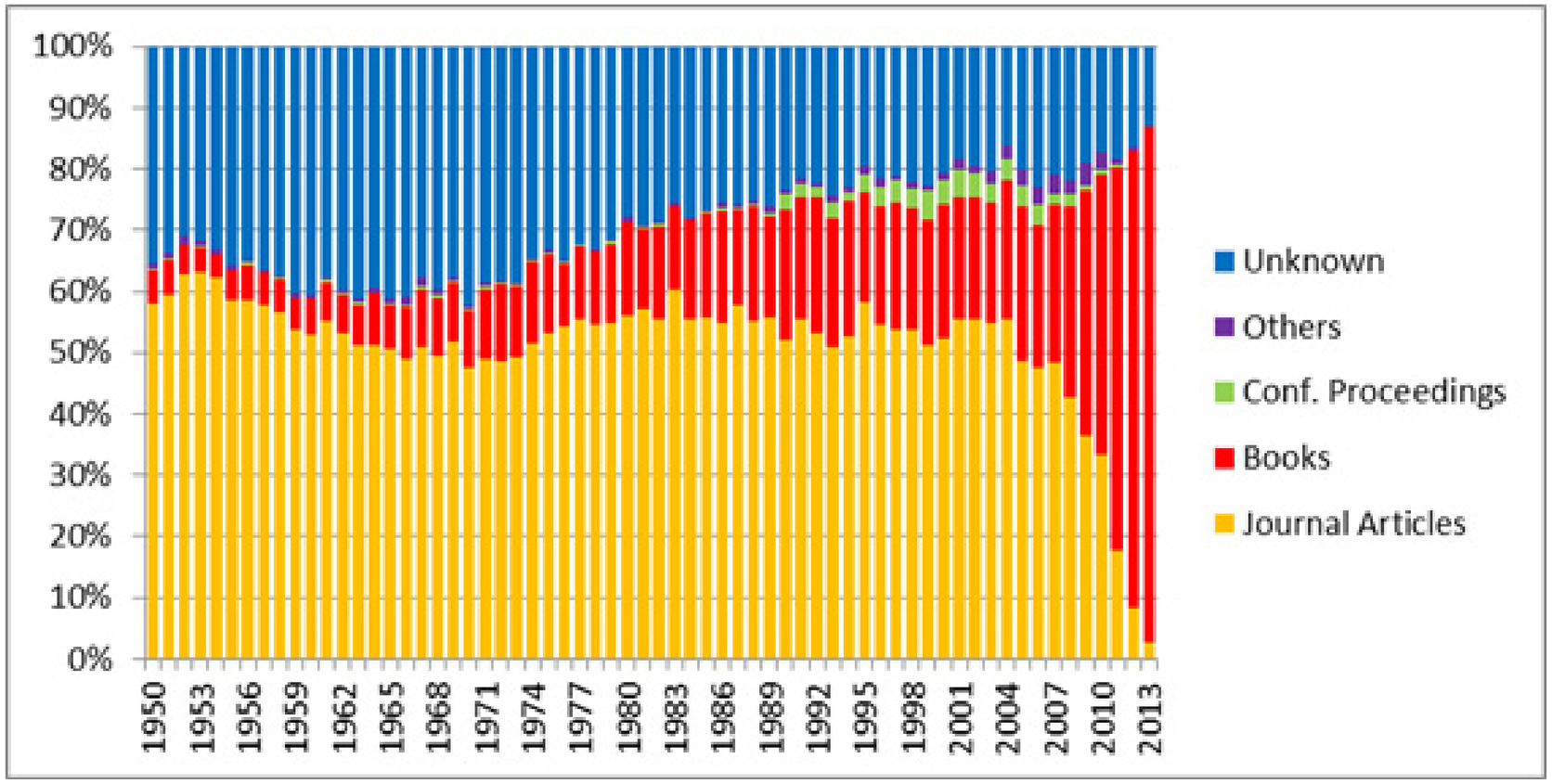
First exploratory analysis:
Analysis of 64,000 documents published in 1950-2013



GOOGLE SCHOLAR AS A SOURCE OF DATA

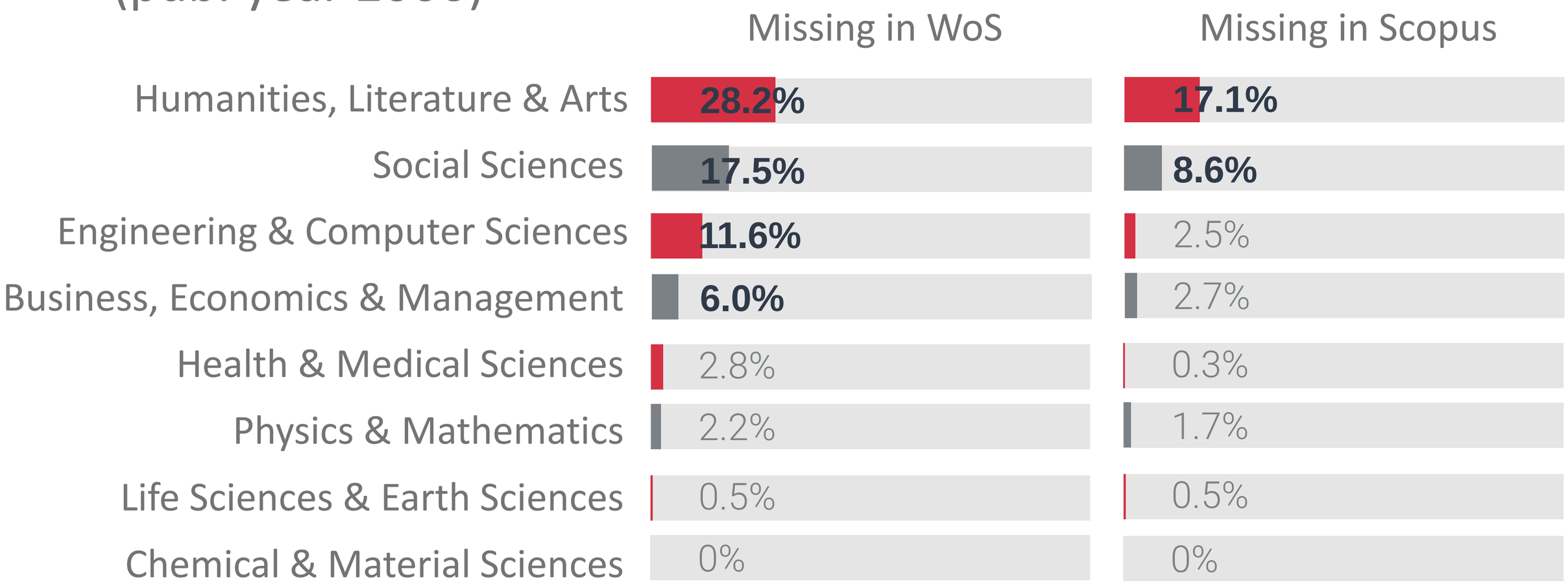
First exploratory analysis:
Analysis of 64,000 documents published in 1950-2013





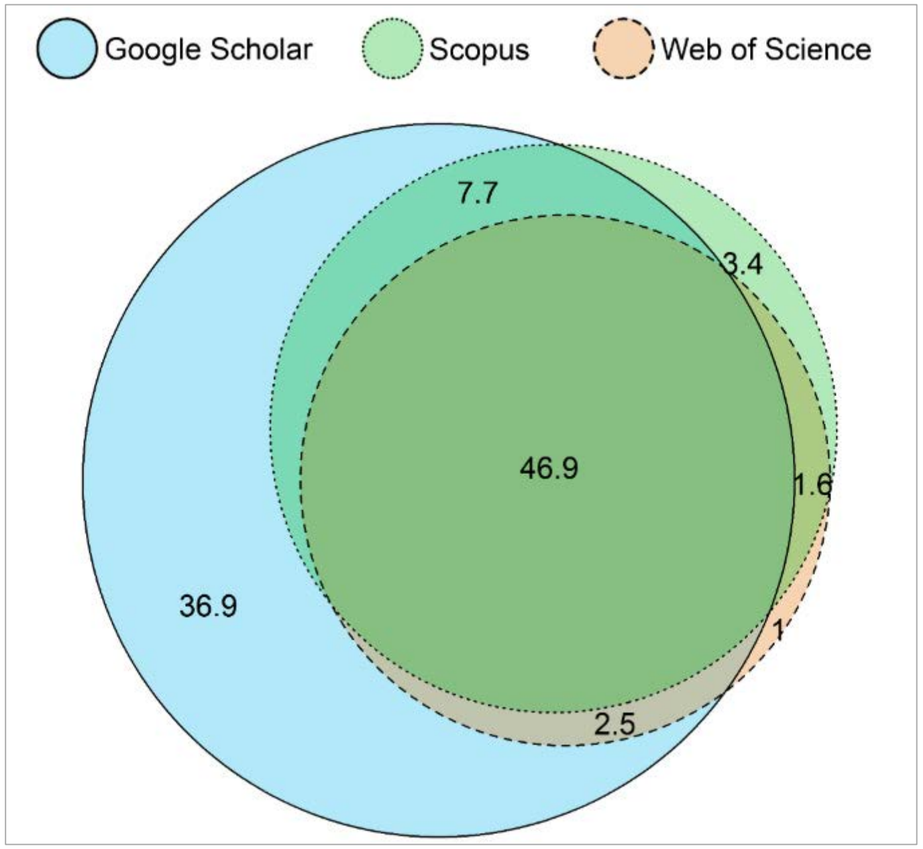
GOOGLE SCHOLAR AS A SOURCE OF DATA

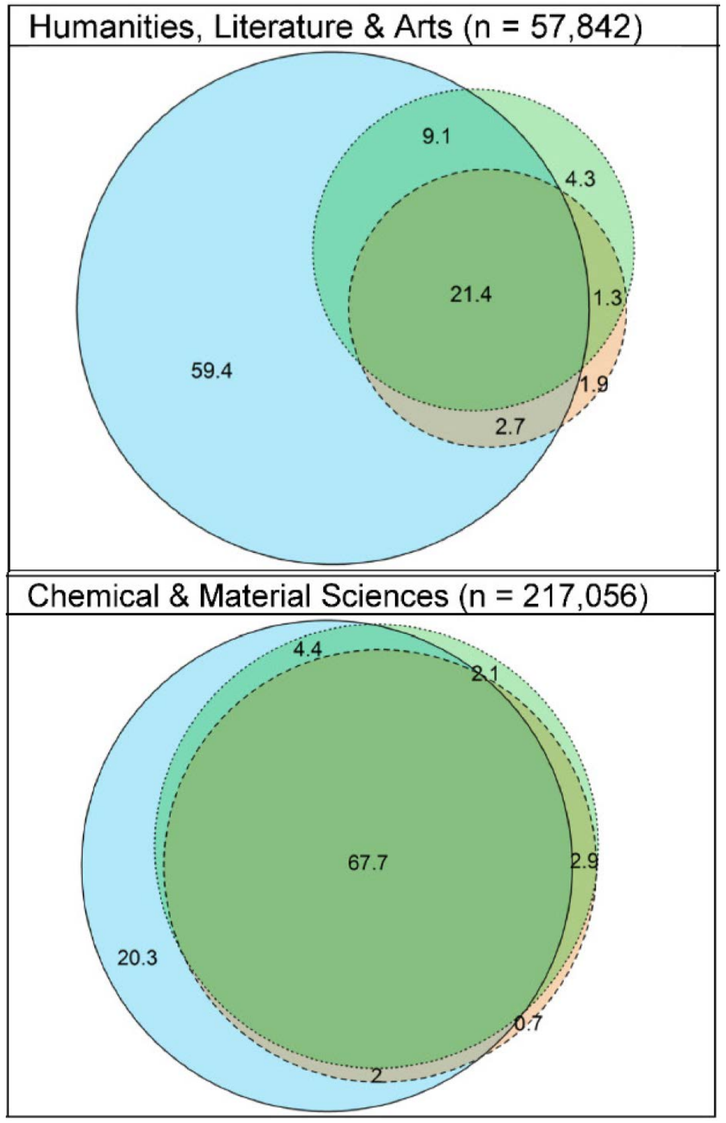
Analysis of highly-cited documents:
Top 10 most cited documents in GS, across 252 subject categories (pub. year 2006)



GOOGLE SCHOLAR AS A SOURCE OF DATA

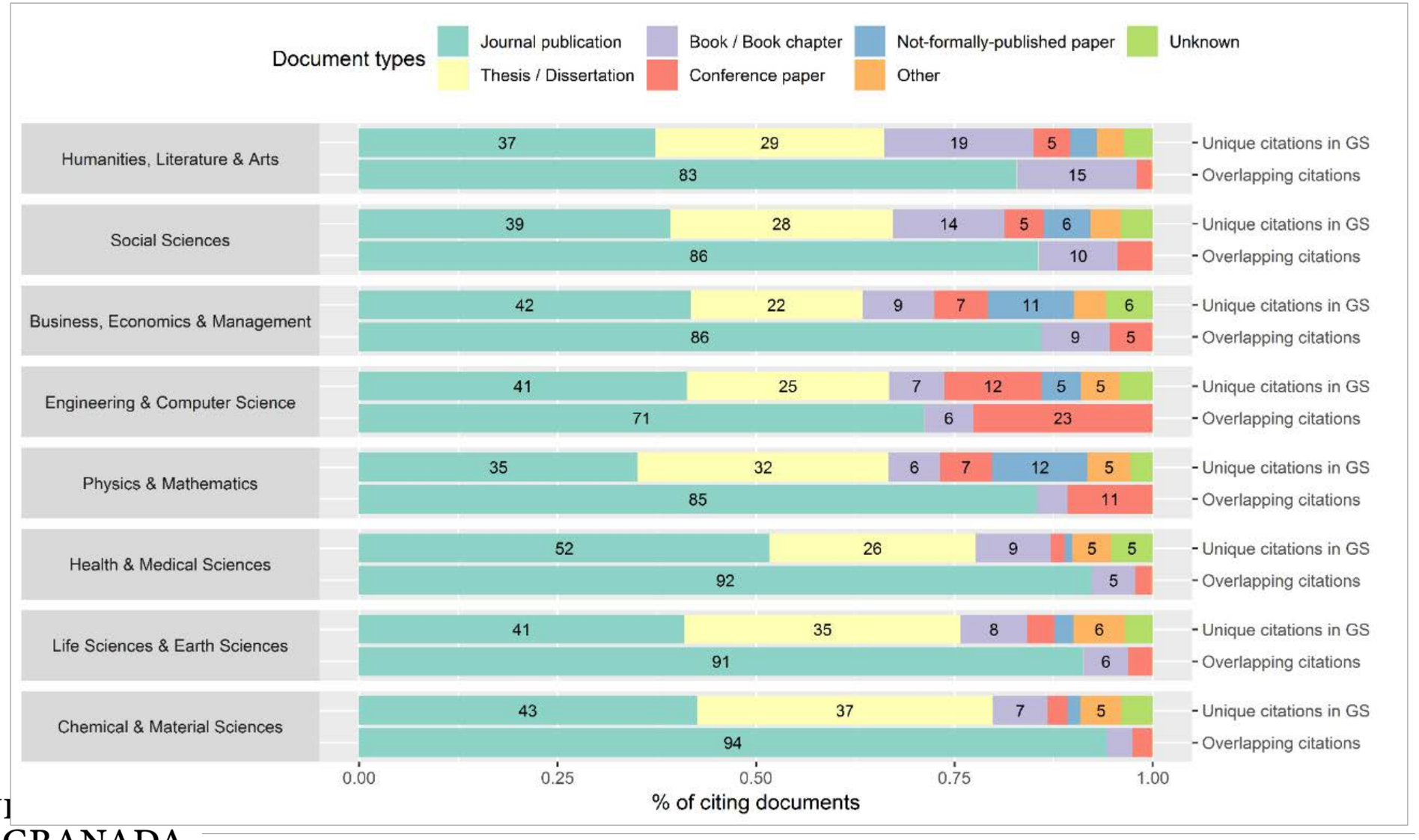
Analysis of citations:
2,448,055 citations to 2,299 highly-cited articles across 252 subject categories





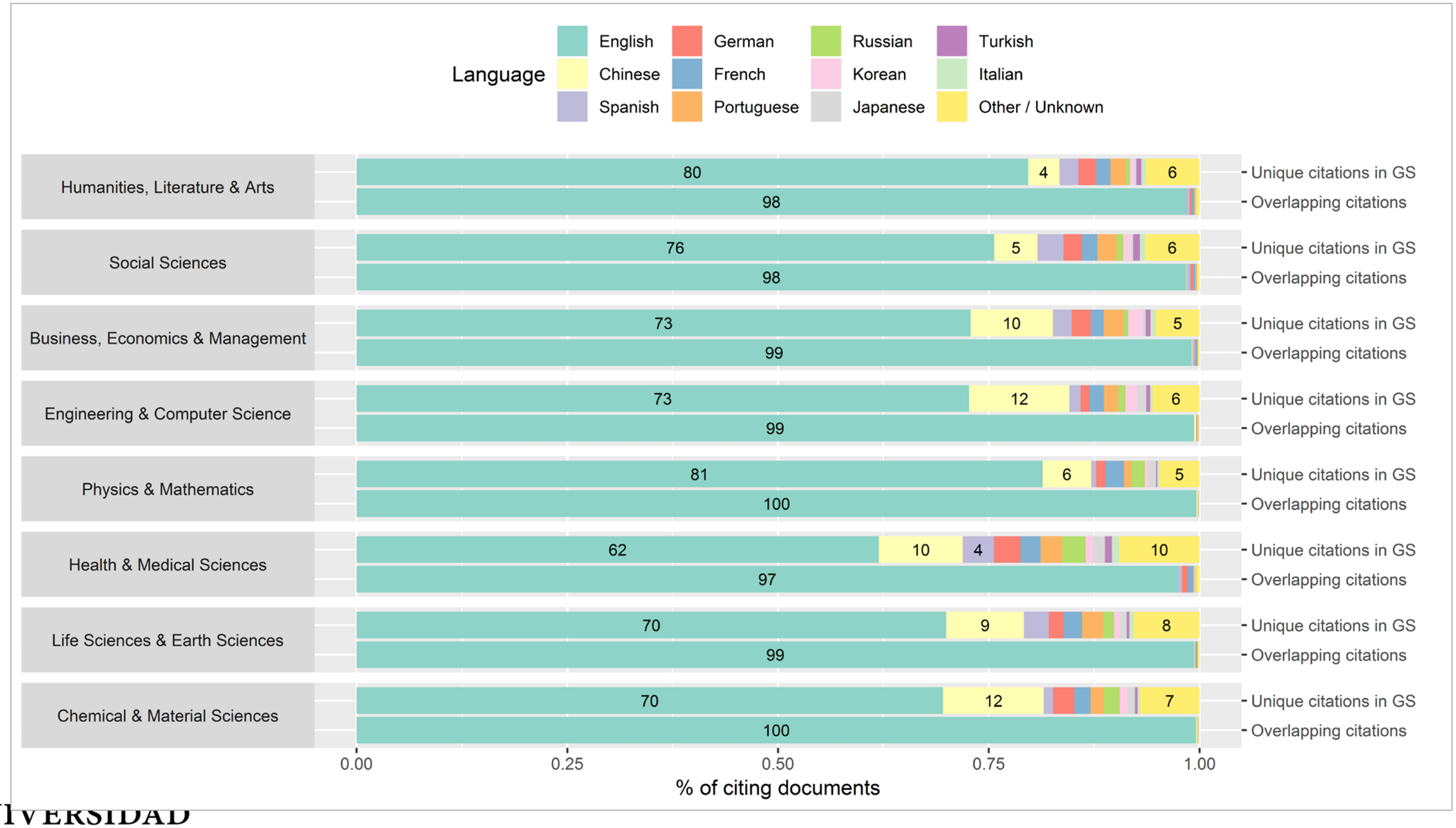
GOOGLE SCHOLAR AS A SOURCE OF DATA

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



GOOGLE SCHOLAR AS A SOURCE OF DATA

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

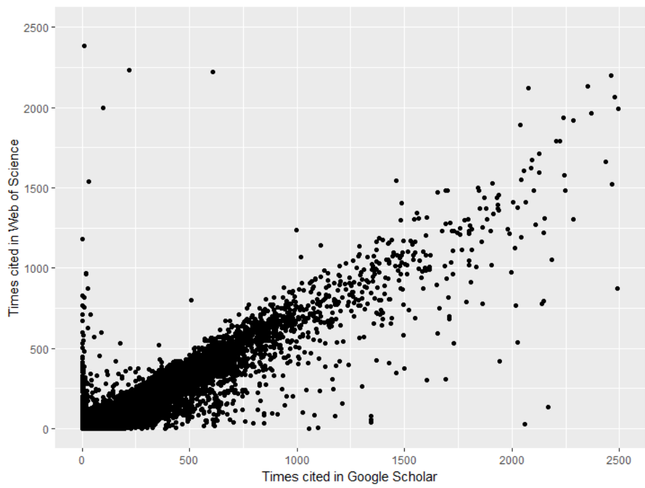


GOOGLE SCHOLAR AS A SOURCE OF DATA

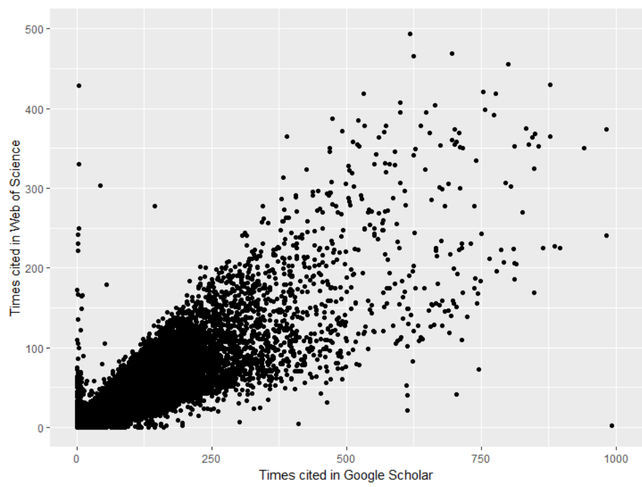
Correlations of citation counts

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**		

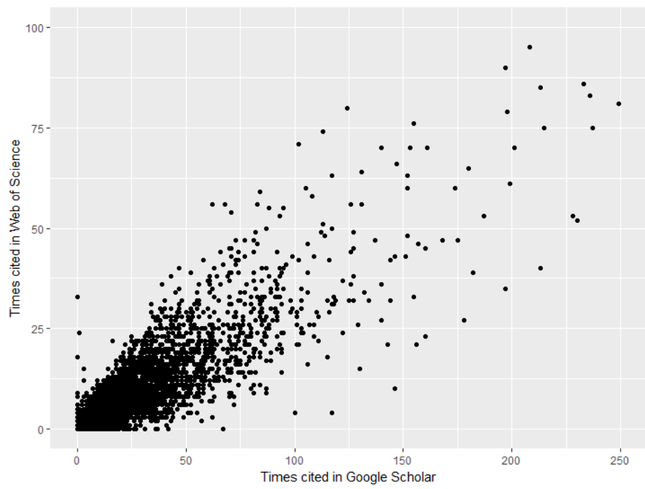
Sciences



Social Sciences



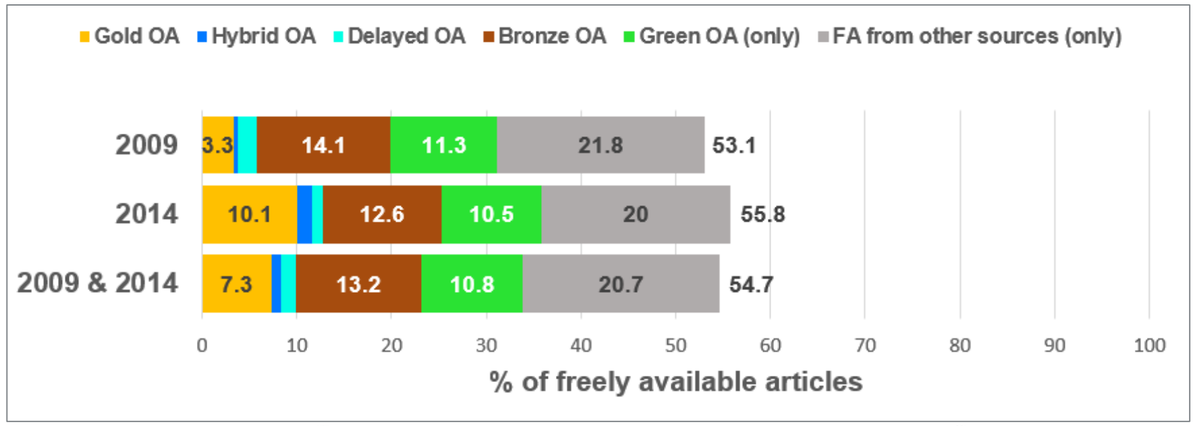
Arts & Humanities

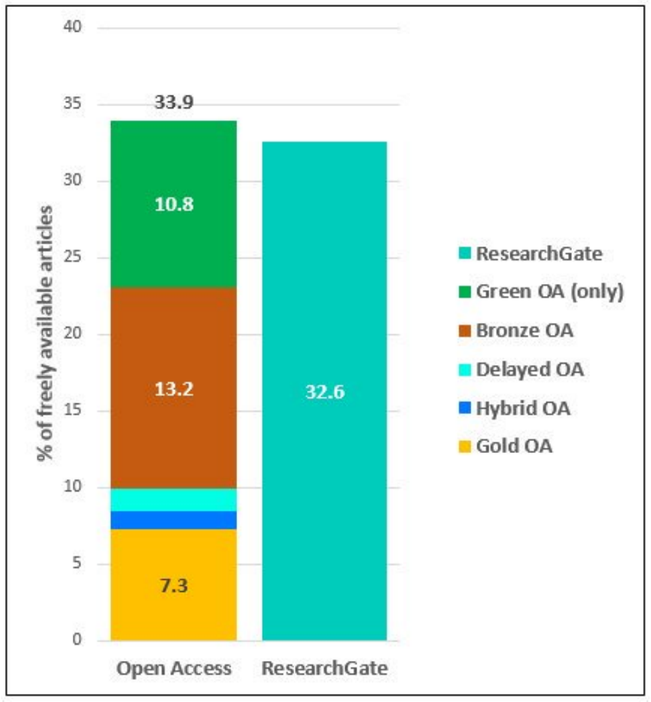


DE GRANADA

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

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):

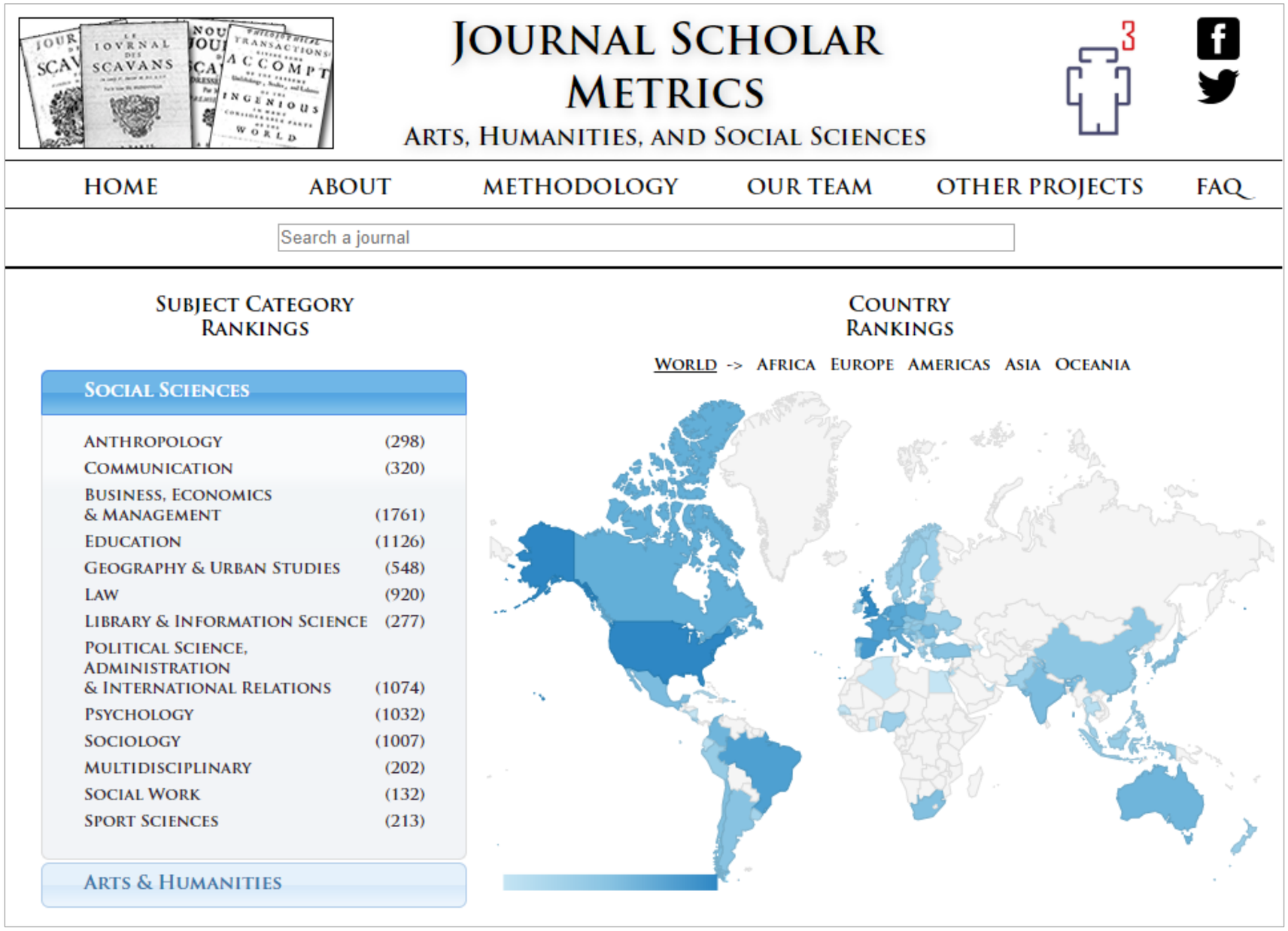
- Duplicate profiles
- Misattributed documents

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%)**

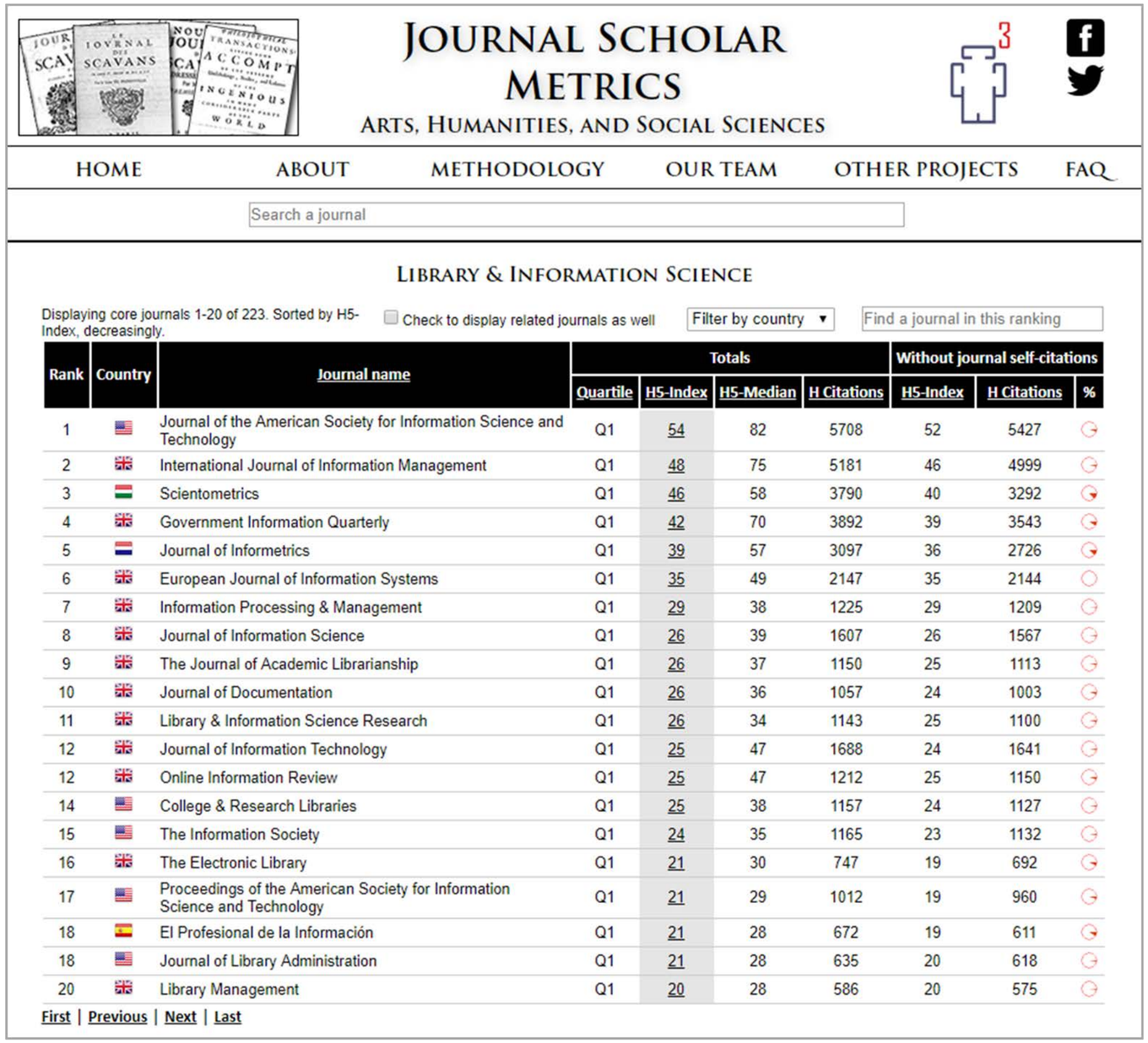


REUSING DATA FROM GOOGLE SCHOLAR

Journal Scholar Metrics

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

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



JOURNAL SCHOLAR METRICS

ARTS, HUMANITIES, AND SOCIAL SCIENCES

HOME | ABOUT | METHODOLOGY | OUR TEAM | OTHER PROJECTS | FAQ

Search a journal

LIBRARY & INFORMATION SCIENCE

Displaying core journals 1-20 of 223. Sorted by H5-Index, decreasingly. Check to display related journals as well. Filter by country. Find a journal in this ranking

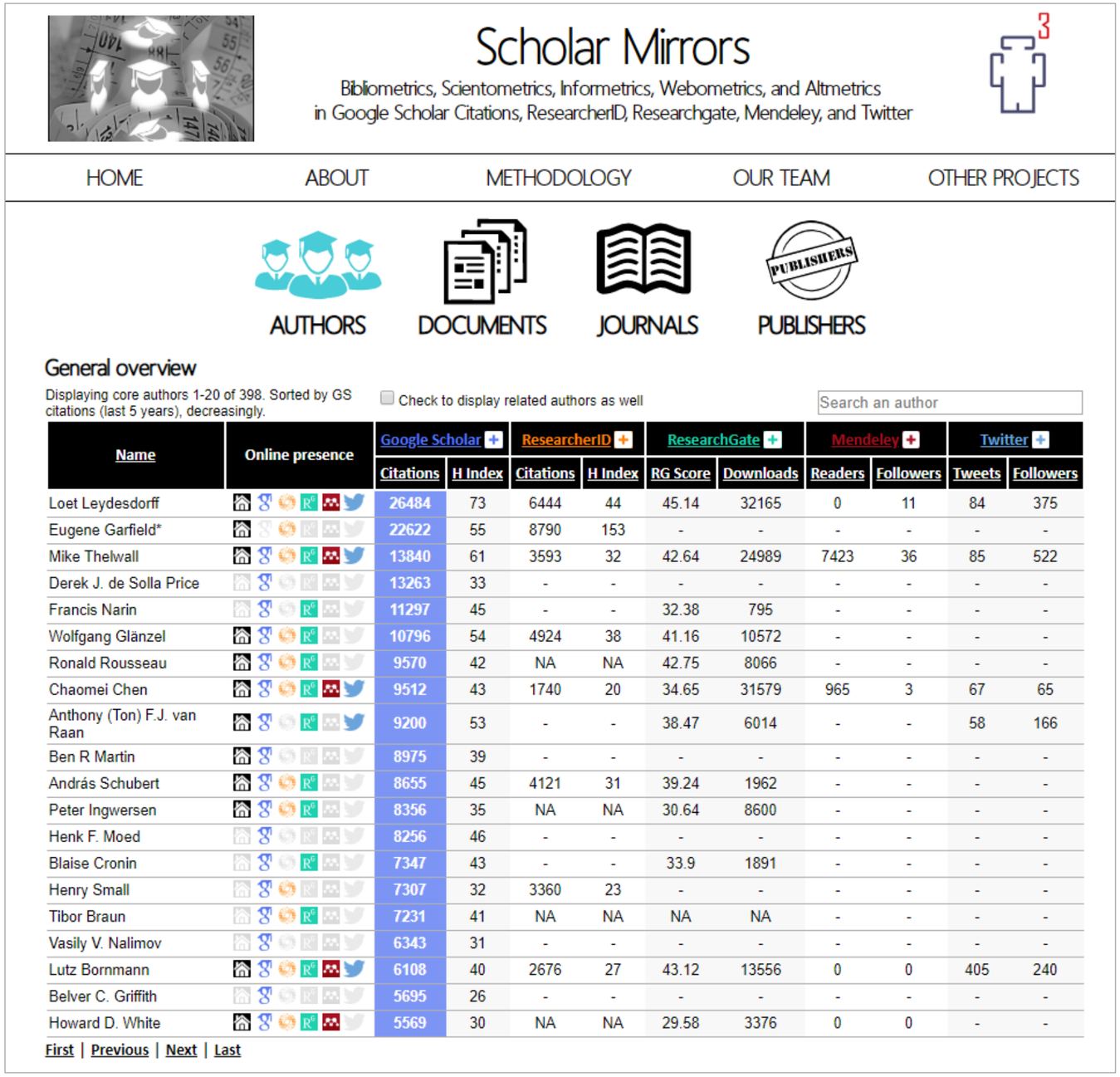
Rank	Country	Journal name	Totals				Without journal self-citations		
			Quartile	H5-Index	H5-Median	H Citations	H5-Index	H Citations	%
1	US	Journal of the American Society for Information Science and Technology	Q1	54	82	5708	52	5427	
2	UK	International Journal of Information Management	Q1	48	75	5181	46	4999	
3	HU	Scientometrics	Q1	46	58	3790	40	3292	
4	UK	Government Information Quarterly	Q1	42	70	3892	39	3543	
5	NL	Journal of Informetrics	Q1	39	57	3097	36	2726	
6	UK	European Journal of Information Systems	Q1	35	49	2147	35	2144	
7	UK	Information Processing & Management	Q1	29	38	1225	29	1209	
8	UK	Journal of Information Science	Q1	26	39	1607	26	1567	
9	UK	The Journal of Academic Librarianship	Q1	26	37	1150	25	1113	
10	UK	Journal of Documentation	Q1	26	36	1057	24	1003	
11	UK	Library & Information Science Research	Q1	26	34	1143	25	1100	
12	UK	Journal of Information Technology	Q1	25	47	1688	24	1641	
12	UK	Online Information Review	Q1	25	47	1212	25	1150	
14	US	College & Research Libraries	Q1	25	38	1157	24	1127	
15	US	The Information Society	Q1	24	35	1165	23	1132	
16	UK	The Electronic Library	Q1	21	30	747	19	692	
17	US	Proceedings of the American Society for Information Science and Technology	Q1	21	29	1012	19	960	
18	ES	El Profesional de la Información	Q1	21	28	672	19	611	
18	US	Journal of Library Administration	Q1	21	28	635	20	618	
20	UK	Library Management	Q1	20	28	586	20	575	

First | Previous | Next | Last

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

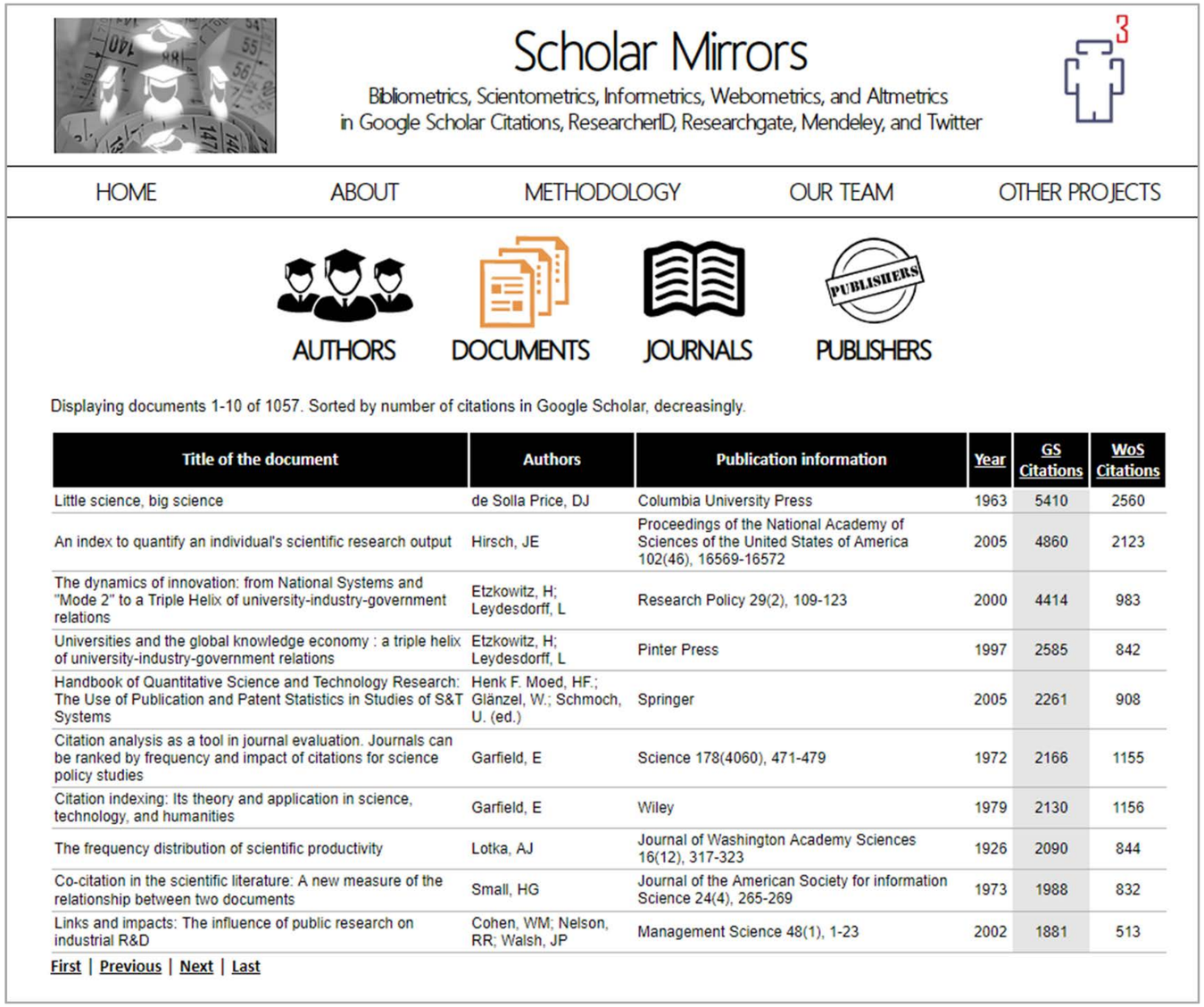
Name	Online presence	Google Scholar		ResearcherID		ResearchGate		Mendeley		Twitter	
		Citations	H Index	Citations	H Index	RG Score	Downloads	Readers	Followers	Tweets	Followers
Loet Leydesdorff		26484	73	6444	44	45.14	32165	0	11	84	375
Eugene Garfield*		22622	55	8790	153	-	-	-	-	-	-
Mike Thelwall		13840	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

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

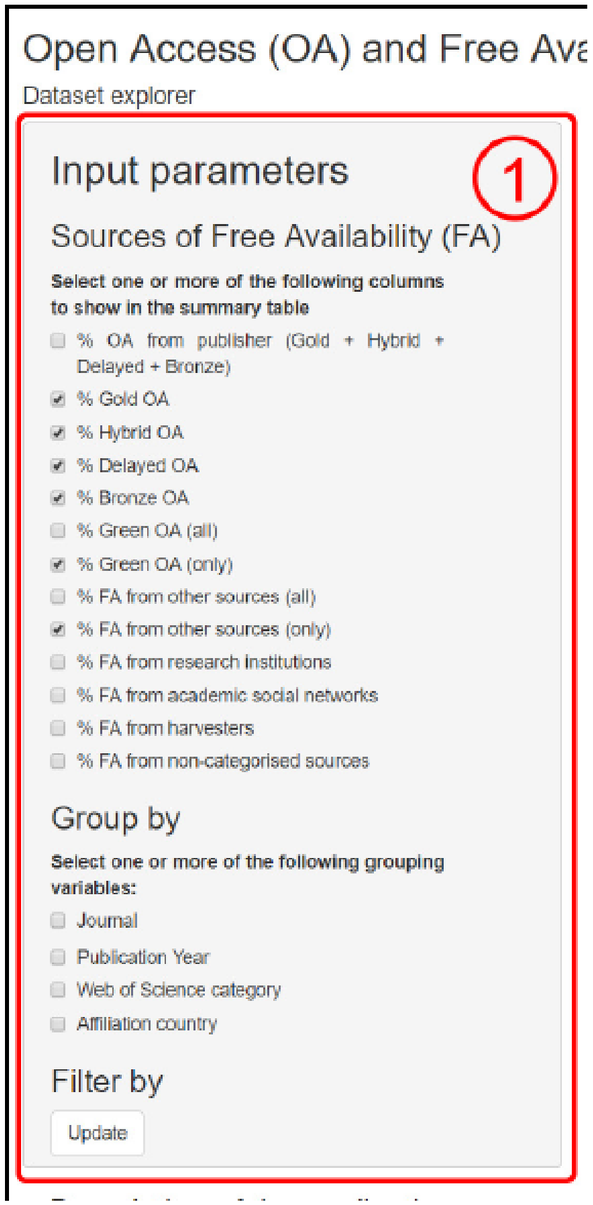
Displaying documents 1-10 of 1057. Sorted by number of citations in Google Scholar, decreasingly.

Title of the document	Authors	Publication information	Year	GS Citations	WoS Citations
Little science, big science	de Solla Price, DJ	Columbia University Press	1963	5410	2560
An index to quantify an individual's scientific research output	Hirsch, JE	Proceedings of the National Academy of Sciences of the United States of America 102(46), 16569-16572	2005	4860	2123
The dynamics of innovation: from National Systems and "Mode 2" to a Triple Helix of university-industry-government relations	Etzkowitz, H; Leydesdorff, L	Research Policy 29(2), 109-123	2000	4414	983
Universities and the global knowledge economy : a triple helix of university-industry-government relations	Etzkowitz, H; Leydesdorff, L	Pinter Press	1997	2585	842
Handbook of Quantitative Science and Technology Research: The Use of Publication and Patent Statistics in Studies of S&T Systems	Henk F. Moed, HF.; Glänzel, W.; Schmoch, U. (ed.)	Springer	2005	2261	908
Citation analysis as a tool in journal evaluation. Journals can be ranked by frequency and impact of citations for science policy studies	Garfield, E	Science 178(4060), 471-479	1972	2166	1155
Citation indexing: Its theory and application in science, technology, and humanities	Garfield, E	Wiley	1979	2130	1156
The frequency distribution of scientific productivity	Lotka, AJ	Journal of Washington Academy Sciences 16(12), 317-323	1926	2090	844
Co-citation in the scientific literature: A new measure of the relationship between two documents	Small, HG	Journal of the American Society for information Science 24(4), 265-269	1973	1988	832
Links and impacts: The influence of public research on industrial R&D	Cohen, WM; Nelson, RR; Walsh, JP	Management Science 48(1), 1-23	2002	1881	513

First | Previous | Next | Last

REUSING DATA FROM GOOGLE SCHOLAR

Open Access dashboard



Results

Download table of DOIs and links to free full texts

Copy URL maintaining current input parameters

Summary table

2

Show 10 entries

Search:

# of documents	% FA from all sources	% Gold OA	% Hybrid OA	% Delayed OA	% Bronze OA	% Green OA (only)	% FA from other sources (only)
All	All	All	All	All	All	All	All
2269022	54.7	7.3	1.1	1.5	13.2	10.8	20.7

Showing 1 to 1 of 1 entries

Previous 1 Next

Download table

Number of freely accessible documents by domain

3

Show 10 entries

Search:

Host	Host type	# of documents	% as only FA provider	# as primary version	% as primary version
All	All	All	All	All	All
www.researchgate.net	social_network	738573	32.7	323372	43.8
europepmc.org	repository	177930	5.1	18312	10.3
www.academia.edu	social_network	166485	4.2	23681	14.1
www.ncbi.nlm.nih.gov	repository	165403	1.8	74109	44.8
citeseerx.ist.psu.edu	harvester	120378	1.8	11203	9.3
arxiv.org	repository	72862	25	72753	99.9
onlinelibrary.wiley.com	publisher	49887	32.8	47712	95.6
www.sciencedirect.com	publisher	47356	26.1	43825	92.5
pdfs.semanticscholar.org	harvester	38164	1	2790	7.3
journals.plos.org	publisher	37984	12.5	37380	98.4

Showing 1 to 10 of 116,271 entries

Previous 1 2 3 4 5 ... 11628 Next

Download table

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

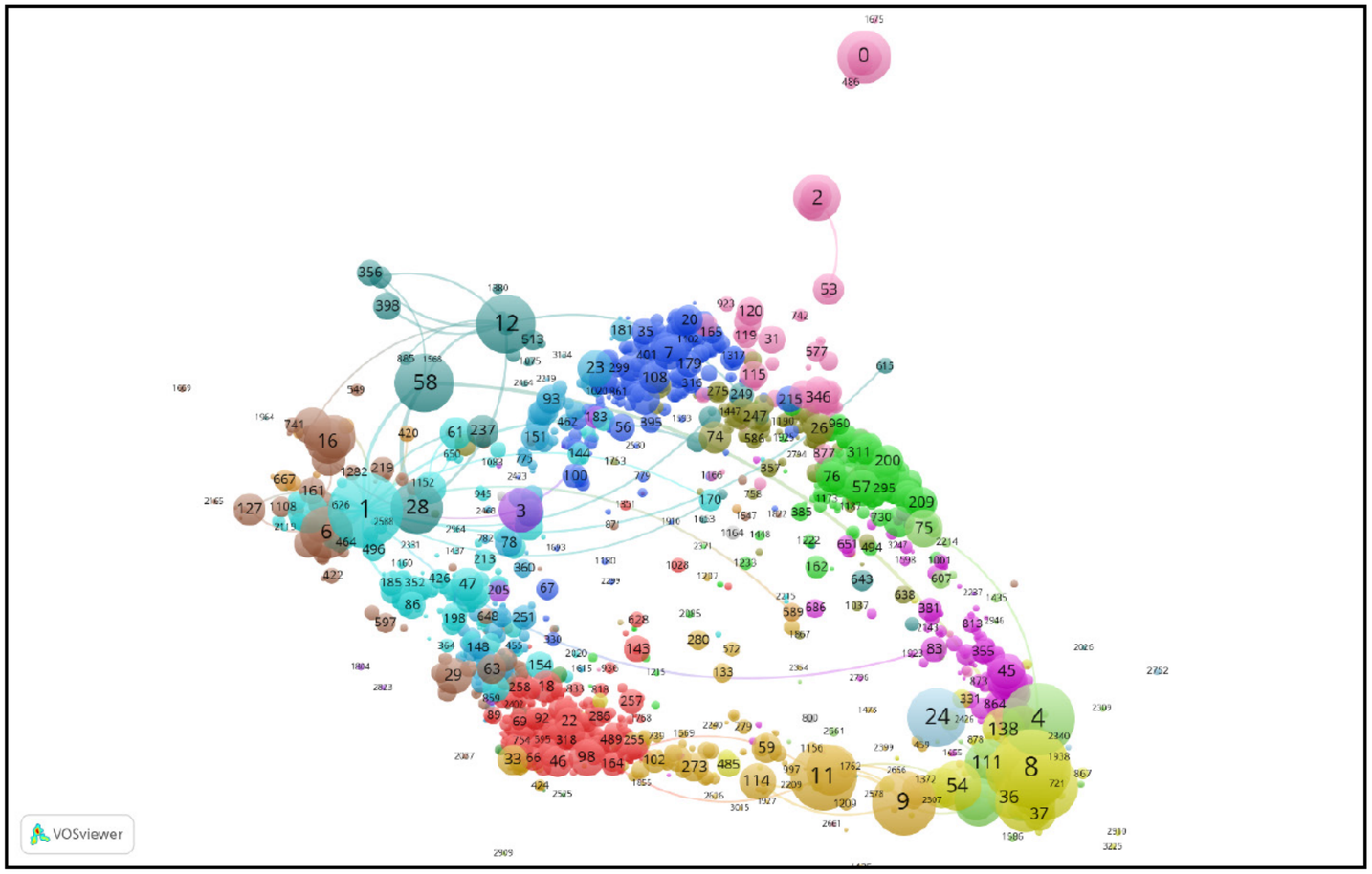


Figure 2. Clusters of documents displayed in the Google Scholar Citations profiles of researchers working in Spain

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