STI 2018 Conference Proceedings

Proceedings of the 23rd International Conference on Science and Technology Indicators

All papers published in this conference proceedings have been peer reviewed through a peer review process administered by the proceedings Editors. Reviews were conducted by expert referees to the professional and scientific standards expected of a conference proceedings.

Chair of the Conference

Paul Wouters

Scientific Editors

Rodrigo Costas
Thomas Franssen
Alfredo Yegros-Yegros

Layout

Andrea Reyes Elizondo
Suze van der Luijt-Jansen

The articles of this collection can be accessed at https://hdl.handle.net/1887/64521

ISBN: 978-90-9031204-0

23rd International Conference on Science and Technology Indicators (STI 2018)
"Science, Technology and Innovation indicators in transition"
12 - 14 September 2018 | Leiden, The Netherlands
#STI18LDN

Classic papers: using Google Scholar to detect the highly-cited documents

Enrique Orduna-Malea[*], Alberto Martín-Martín[**] and Emilio Delgado López-Cózar[**]

[*] *enorma@upv.es*
Universitat Politècnica de València, Camino de Vera s/n, 46022 (Spain)
[* *] *albertomartin@ugr.es; edelgado@ugr.es*
Facultad de Comunicación y Documentación, Universidad de Granada, Colegio Máximo de Cartuja s/n, 18071 (Spain)

Abstract
In June 2017 *Google Scholar* launched a new product called *Classic Papers*. This service currently displays the most cited English-language original research articles by fields and published in 2006. The main goal of this work is to describe the foremost features of this new service, as well as to highlight its main strengths and weaknesses. To do this, a total of 2,515 records were extracted. For each record, bibliographic data (broad subject category and subcategory; Title of the document; URL; Authors, *Google Scholar Citation* profiles' URL; Citations received) were gathered. It is finally concluded that, although the product is easy to use and provides original data about highly cited documents at the level of disciplines, it still suffers of some methodological concerns (related to the subject classification of documents and the use of homogenous visualization threshold regardless the discipline) that jeopardizes the utility of this product for bibliometric purposes.

Origins of the *Citations Classics*

In 1969 Garfield had already compiled a list of the top 50 most cited articles published in 1967 (Figure 1). In that list he already used the term "classics" to refer to those highly cited documents. Six years later he prepared a similar list, but this time about articles published between 1961 and 1972. This list comprised the top 50 most cited articles published in that period, and he again used the term "classics" to refer to those works.

Figure 1. Most cited articles published in 1967 (Garfield, 1971).

RANK	TOTAL TIMES CITED	AUTHOR	JOURNAL	VOL	PAGE	YEAR
1	2363	LOWRY OH	J BIOL CHEM	193	265	51
2	654	REYNOLDS ES	J CELL BIOL	17	208	63
3	561	LUFT JH	J BIOPHYS BIOCHEM CY	9	409	61
4	518	FISKE CH	J BIOL CHEM	66	375	25
5	457	FOLCH J	J BIOL CHEM	226	497	57
6	456	BRAY GA	ANAL BIOCHEM	1	279	60
7	388	SABATINI DD	J CELL BIOL	17	19	63
8	381	SPACKMAN DH	ANAL CHEM	30	1190	58
9	364	GORNALL AG	J BIOL CHEM	177	751	49
10	333	LINEWEAVER H	J AMER CHEM SOC	56	658	34
11	286	BURTON K	BIOCHEM J	62	315	56
12	275	DUNCAN DB	BIOMETRICS	11	1	55
13	274	SCHEIDEGGER JJ	INT ARCH ALLERGY APP	7	103	55
14	241	DOLE VP	J CLIN INVEST	35	150	56
15	226	DAVIS BJ	ANN NY ACAD SCI	121	404	64
16	223	NELSON N	J BIOL CHEM	153	375	44
17	223	REED LJ	AMER J HYG	27	493	38
18	218	MOORHEAD PS	EXP CELL RES	20	613	60
19	217	MARMUR J	J MOL BIOL	3	208	61
20	207	JACOB F	J MOL BIOL	3	318	61
21	203	WATSON ML	J BIOPHYS BIOCHEM CY	4	475	58
22	197	PALADE GE	J EXP MED	95	285	52
23	182	KARNOVSKY MJ	J BIOPHYS BIOCHEM CY	11	729	61
24	182	MARTIN RG	J BIOL CHEM	236	1372	61
25	175	SMITHIES O	BIOCHEM J	61	629	55
26	163	BARTLETT GR	J BIOL CHEM	234	466	59
27	162	BARKER SB	J BIOL CHEM	138	535	41
28	160	EAGLE H	SCIENCE	130	432	59
29	156	ROSENFELD AH	REV MOD PHYS	39	1	67
30	156	GELLMANN M	PHYS REV	125	1067	62
31	153	TREVELYAN WE	NATURE LOND	166	444	50
32	140	WARREN L	J BIOL CHEM	234	1971	59
33	140	ANDREWS P	BIOCHEM J	91	222	64
34	139	MONOD J	J MOL BIOL	12	88	65
35	136	SCHMIDT G	J BIOL CHEM	161	83	45
36	134	BARDEEN J	PHYS REV	108	1175	57
37	134	DEDUVE C	BIOCHEM J	60	604	55
38	131	KARPLUS M	J CHEM PHYS	30	11	59
39	131	AHLQUIST RP	AM J PHYSIOL	153	586	48
40	130	DUBOIS M	ANAL CHEM	28	350	56
41	128	ELLMAN GL	ARCH BIOCHEM BIOPHYS	82	70	59
42	125	WARBURG O	BIOCHEM Z	310	384	41
43	125	GELLMANN M	PHYSICS	1	63	64
44	124	MANDELL JD	ANAL BIOCHEM	1	66	60
45	123	DOLE VP	J BIOL CHEM	235	2595	60
46	122	LITCHFIELD JT	J PHARMAC EXP THER	96	99	49
47	122	MILLONIG G	J APPL PHYSICS	32	1637	61
48	118	FRIEDEMANN TE	J BIOL CHEM	147	415	43
49	119	MOORE S	J BIOL CHEM	211	907	54
50	119	JAFFE HH	CHEM REV	53	191	53

Garfield revisited this topic repeatedly in the following years. No less than 17 essays about the "citation classics"[1] of various scientific fields or journals were published, and some of them stimulated a discussion on the meaning and influence of this kind of studies (immortality, obliteration, productivity, genre, Nobel prizes…). Other essays (more than 80) were dedicated to examining the most cited papers, books, and authors in various disciplines, specialties, journals, or countries. On top of this foundation, *Thomson Scientific* first, *Thomson Reuters* later, and *Clarivate Analytics* today, built the *Essential Science Indicators* (ESI), which every year presents the most cited documents of the last decade.

While the use of highly-cited documents in research evaluation has been studied, the conditions that determine whether a document can be considered highly-cited are not yet globally agreed (Bornmann, 2014).

Google Scholar's Classic Papers

The appearance of *Google Scholar* opened up new possibilities in this field. Its birth in 2004 signalled a revolution in the way scientific publications were searched, retrieved and accessed (Orduna-Malea, Martín-Martín, Ayllón, & Delgado López-Cózar, 2016). The capacity of *Google Scholar* to identify highly-cited documents has been already treated in the literature (Martín-Martín, Orduna-Malea, Harzing, & Delgado López-Cózar, 2017).

Since June 2017, *Google* started providing a new service called *Classic Papers* (GSCP), which contains lists of highly-cited documents by discipline: the top 10 most cited English-language original research articles published in 2006 in 252 subject categories, according to the data available in *Google Scholar* as of May 2017. In July of 2018 Google Scholar *Metrics* was updated, but a new version of *Classic Papers* was not released. Furthermore, the link to

[1] All of them available at http://garfield.library.upenn.edu/citationclassicessays.html

the 2017 edition of *Classic Papers* was removed from the interface, although the product is still accessible[2].

The criteria used by this product to include highly-cited documents are the following:

- They must have been published in 2006
- They must be journal articles, articles deposited in repositories, or conference communications.
- They must describe original research. Review articles, introductory articles, editorials, guides, commentaries, etc. are explicitly excluded.
- They must be written in English.
- They must be among the top 10 most cited documents in their respective subject category.
- They must have received at least 20 citations.

The goal of this study is to assess this new product in order to gauge its reliability and validity for identifying highly-cited documents, and to find its main strengths and weaknesses.

Methods

We first extracted all the information available in *GSCP*. For this purpose, a custom script was developed which scraped all the relevant information, and saved it as a table in a spreadsheet file. The information extracted was:

- Broad subject categories and subcategories.
- Bibliographic information of the documents, including:
  - Title of the document, and URL pointing to the corresponding *Google Scholar* record.
  - Authors (including URL to *Google Scholar Citations* profile when available), name of the publication venue, and year of publication.
  - Name and URL to *Google Scholar Citations* profile of showcased author.
  - Number of citations the document had received (as of May 2017).

A total of 2,515 records were retrieved by July 2017.

Results

Data visualization

Articles are classified in 294 subject categories, which in turn are grouped in eight broad scientific areas (Table 1). Since there are 42 subject categories appearing in two broad scientific areas, there are 252 unique subject categories.

[2] https://scholar.google.com/citations?view_op=list_classic_articles&hl=en&by=2006

Table 1. Number of subject categories in each broad scientific area in *GSCP*.

Areas	Number of subject categories
Health & Medical Sciences	68
Engineering & Computer Science	57
Social Sciences	51
Life Sciences & Earth Sciences	38
Humanities, Literature & Arts	25
Physics & Mathematics	23
Chemical & Material Sciences	17
Business, Economics & Management	15

Each of these 252 categories presents 10 articles, except French Studies, which only has 5 with at least 20 citations, which is the self-imposed minimum used by *Google Scholar*. That is the reason why the total number of articles is 2,515 instead of 2,520 (252 times 10).

One of the innovative aspects of the product is that it displays the link to the *Google Scholar Citations* profile of some of the authors of the article. 31% of the articles (654) displayed in *GSCP* lack such a link, and there are significant differences among disciplines. For example, in 'Chemical & Material Sciences', 5 out of the 17 subdisciplines considered (0.29%) display links to author profiles for all documents included in the subdiscipline, whereas in 'Humanities, Literature & Arts', in none of the 25 subcategories can we find at least one author with a public profile for each of the 10 documents (Table 2).

Table 2. Subcategories with at least one document whose author is linked to an author profile

Category	Subcategories	SWP	%
Life Sciences & Earth Sciences	38	7	0,18
Business, Economics & Management	15	4	0,27
Chemical & Material Sciences	17	5	0,29
Engineering & Computer Science	57	15	0,26
Humanities, Literature & Arts	25	0	0,00
Health & Medical Sciences	68	6	0,09
Physics & Mathematics	23	3	0,13
Social Sciences	51	5	0,10
TOTAL	294	45	

Table 3 shows the subcategories in which there is a higher number of highly-cited documents for which no author profile is available. As we can observe, 'American Literature & Studies' and, unexpectedly, 'Plastic & Reconstructive Surgery', are at the top of this list.

Table 3. Subcategories in *GSCP* in which most of the documents are written by authors that haven't set up a public *Google Scholar Citations* profile.

Subcategories	Number of papers for which no author has a public GSC profile
American Literature & Studies	9
Plastic & Reconstructive Surgery	9
Drama & Theater Arts	8
International Law	8
African Studies & History	7
Dentistry	7
Ethnic & Cultural Studies	7
Literature & Writing	7
Visual Arts	7

Most of the articles displayed in *GSCP* are written in collaboration by several co-authors, and even if more than one has a public *Google Scholar Citations* profile, only one is prominently displayed in the record. The system seems to give preference to the first author, then to the last author, and if neither of these have a profile, it selects whatever profile is available first according to author order.

Reliability and validity

There are four critical aspects about which we should know more precise information.

1) What does *GSCP* understands as a research article?

Although they declare that they are "…articles that presented new research"[3], we ask: how have they identified research articles from those that are not research articles? What constitutes an introductory article and how have they identified them? What do they mean when they add a disconcerting "etc." when they list the excluded document types? "Etc." is rarely admissible in Science, where all explanations should be precise. This issue is important because it may be the case that some articles that don't meet these requisites have been included, or the opposite, that some articles that do meet the requisites are missing.

It is important to remember that defining the typology of a document is not an easy task, and that even traditional bibliographic databases like *Web of Science* or *Scopus* have not been able to solve this issue completely. There are many discrepancies in how each of these databases defines the typology of the documents they cover. This happens frequently with review articles. There are also abundant internal inconsistencies in the databases.

2) Subject classification of the articles

This task involves assigning each article to one of 252 subject categories, and it is a crucial issue for the correct development of the product, but also very thorny. There are two fundamental questions we may ask regarding this issue:

a) Which criteria have they adopted to carry out the subject classification?

It seems clear that the classification scheme they have selected is the same they use in *Google Scholar Metrics*, their annual ranking of scientific journals. The only difference is the elimination of eight subject categories ('Physics & Mathematics'; 'Business,

[3] https://scholar.googleblog.com/2017/06/classic-papers-articles-that-have-stood.html

Economics & Management'; 'Chemical & Material Sciences'; 'Health & Medical Sciences'; 'Engineering & Computer Science'; 'Life Sciences & Earth Sciences'; 'Social Sciences'; 'Humanities, Literature & Arts') which are referred to as "general", because their title is the same as the broad scientific area where they are included.

At first, the elimination of these categories should not pose any problem, because the journals included in those categories are also classified in other subject categories (sometimes up to four other). However, there are journals which are only classified in these generic categories. Have the articles published in these journals been classified in other subject categories?

We have checked that articles published in multidisciplinary journals (such as *Nature*, *Science*, or *PNAS*) have been indeed classified *ad hoc* in their respective subject categories according to the topic of the articles. It seems that the articles published in journals with a broad scope have also been classified in the correct subject categories (*Journal of the American Chemical Society*, *IEEE Transactions on Industrial Electronics*, *The New England Journal of Medicine*, *JAMA*, *The Lancet*, *Qualitative Inquiry*, *Scientific Reports*, *PLoS Biology*, *Reviews of Modern Physics*, *Procedia-Social and Behavioral Sciences*).

b) How have they classified the articles published in multidisciplinary journals and journals with a broad scope?

Most services rely on journal-level classifications instead of article-level classifications. Recently *Dimensions* database started classifying at the level of contributions with some inconsistencies detected (Orduna-Malea & Delgado López-Cózar, 2018). In this sense, how has *Google Scholar* solved this problem? In most cases articles are simply assigned to the same categories where the journal has been classified, without paying attention to the actual topic of the article.

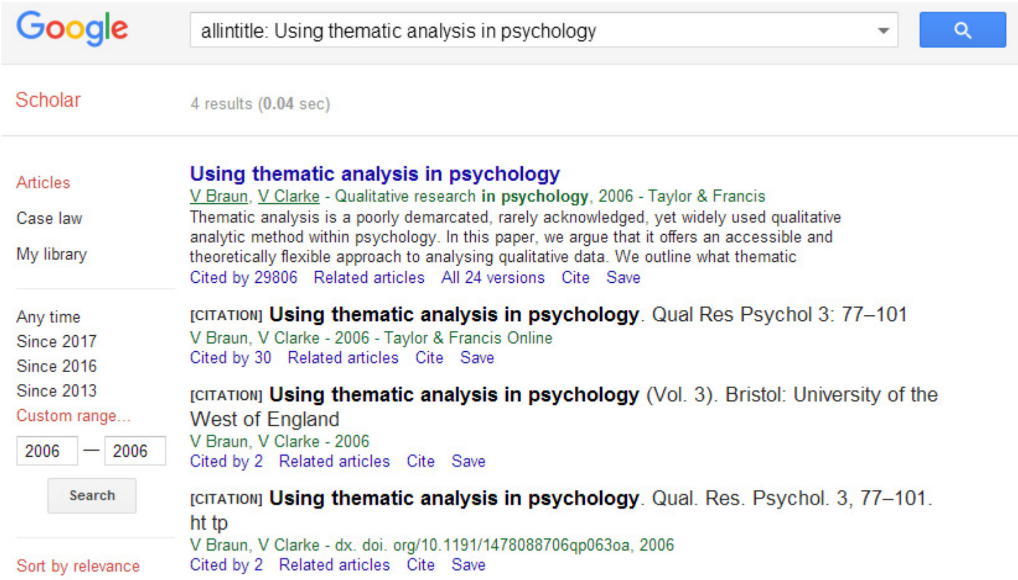
This approach, the most commonly used in bibliometrics, is ill-suited for multidisciplinary journals and the other journals with a broad scope that are published in most disciplines. It is known that the ESI classifies multidisciplinary articles according to the subject categories of the journals publishing the articles that cite them as well as to the journals of the articles cited by them, an incontrovertible approach.

3) <u>How does *Google Scholar* handle document versions?</u>

Can we be sure they have successfully merged together all the versions indexed in *Google Scholar* of these documents? Otherwise, the citation counts of the documents might be scattered in several records.

Previous studies have shown that this is an important issue when we are talking about highly-cited articles (Martín-Martín, Ayllón, Delgado López-Cózar, & Orduna-Malea, 2015). It seems, as Figure 2 evidences, that there are still some records that refer to the same highly-cited documents that appear in *GSCP* which haven't been merged with the main record (the one with the most citations).

Figure 2. Example of versions that have not been properly merged to the main record.



d) <u>What is the threshold selected to visualize a “classic paper”?</u>

Why did they decide to set this number to 10 articles in each subject category? Why is this threshold the same for the 252 subject categories?

This decision goes against logic and long-established bibliometric practices, where the different natures of the various scientific disciplines have long been acknowledged. Different scientific communities have different citation habits and different sizes in terms of number of researchers. In order to illustrate this inconsistency, the 10 WoS categories with the highest number of papers published in 2006, and the 10 categories with the lowest number of papers published in the same year are displayed in Table 4. Next to the number of papers, another column shows the fraction that 10 articles is respect to the total amount of articles in the category.

Table 4. Number of papers classified in the 10 most productive (top) and least productive (down) WoS categories

Web of Science Categories	N papers	% covered by 10 documents
Engineering Electrical Electronic	86,568	0.012
Computer Science Artificial Intelligence	61,137	0.016
Materials Science Multidisciplinary	53,671	0.019
Physics Applied	49,267	0.020
Biochemistry Molecular Biology	47,259	0.021
Chemistry Physical	39,715	0.025
Telecommunications	37,641	0.027
Computer Science Theory Methods	36,233	0.028
Optics	33,660	0.030
Physics Condensed Matter	32,806	0.030

Web of Science Categories	N papers	% covered by 10 documents
Psychology Mathematical	498	2.008
Primary Health Care	484	2.066
Medical Ethics	474	2.110
Dance	401	2.494
Literature American	399	2.506
Andrology	378	2.646
Poetry	368	2.717
Literature Slavic	254	3.937
Folklore	205	4.878
Literature African Australian Canadian	175	5.714

While in ‘Engineering Electrical Electronic’ and ‘Computer Science Artificial Intelligence’ those 10 documents make up barely 0.01% of the total, in ‘Folklore’ and ‘Literature African Australian Canadian’, 10 articles make up more than 5% of the articles in the category.

This productive disparity among disciplines goes together with also huge differences in citation patterns. The maximum and minimum number of citations in the 10 articles displayed in *GSCP* in the 10 categories with highest (top) and lowest (down) number of citations is shown in Table 5. This way it is easy to see the problem of selecting the same citation threshold (20) for all subject categories.

Table 5. Citations in the 10 subject categories in *GSCP* with highest (top) and lowest (down) numbers of citations overall.

Subcategories	Citations (10 most cited articles)		
	Maximun	Minimum	Total
Information Theory	18,648	1,179	51,987
Psychology	29,294	1,181	42,226
Cell Biology	17,121	1,278	36,359
Oncology	6,987	2,411	35,763
Bioinformatics & Computational Biology	9,981	1,555	34,680
Condensed Matter Physics & Semiconductors	8,415	1,640	34,379
Immunology	5,706	1,706	23,200
Economics	3,112	1,883	23,048
Molecular Modeling	9,745	766	22,823
Astronomy & Astrophysics	6,624	1,056	21,854

Subcategories	Citations (10 most cited articles)		
	Maximun	Minimum	Total
Literature & Writing	353	72	1,263
Visual Arts	155	89	1,101
Film	536	37	1,049
Technology Law	75	41	1,014
European Law	178	63	978
Middle Eastern & Islamic Studies	225	58	966
Canadian Studies & History	182	42	706
American Literature & Studies	81	32	545
Drama & Theater Arts	69	34	450
French Studies	32	20	131

Garfield acknowledges this problem when discussing what a "citation classic" is. He said "Citation rates differ for each discipline [...] In general, a publication cited more than 400 times should be considered a classic; but in some fields with fewer researchers, 100 citations might qualify a work"[4]. The highly cited papers available in the ESI follows the same principles delineated by Garfield. Today the product "lists the top cited papers over the last 10 years in 22 scientific fields. Rankings are based on meeting a threshold of the top 1% by field and year based on total citations received"[5]

Conclusions

The main advantage of *GSCP* is the simplicity of the product (a list of the most cited articles in each discipline, with a simple browsing interface). It is organized by broad scientific areas and inside of them by subject categories. Three clicks are enough to reach the documents or the public *Google Scholar Citations* profiles of their authors. Only minimal information is offered. As a whole, the product displays just over 2,500 highly cited articles. Each article presents the most basic bibliographic information.

However, despite the product is easy to use and provides original data about highly cited documents per discipline, it still suffers of some methodological concerns, mainly related to the subject classification of documents and the use of homogenous visualization threshold regardless the discipline, that jeopardizes the utility of this product for bibliometric purposes.

[4] Garfield, E. Short History of Citation Classics Commentaries. Available at http://garfield.library.upenn.edu/classics.html

[5] https://images.webofknowledge.com/images/help/WOS/hs_citation_applications.html

In addition to this, the lack of transparency constitutes a methodological concern, since *Google Scholar* does not declare in detail how the product has been developed.

Acknowledgements
Alberto Martín-Martín enjoys a four-year doctoral fellowship (FPU2013/05863) granted by the *Ministerio de Educación, Cultura, y Deportes* (Spain).

References
Bornmann, L. (2014). How are excellent (highly cited) papers defined in bibliometrics? A quantitative analysis of the literature. *Research Evaluation*, 23(2), 166-173.

Garfield, E. (1971). Citation indexing, historio-bibliography and the sociology of science. *Current Contents*, 6, 156-157.

Garfield, E. (1974). Selecting the All-Time Citation Classics. Here Are the Fifty Most Cited Papers for 1961-1972. *Current Contents*, 2, 5-8.

Garfield, E. (1977). Introducing Citation Classics: The human side of scientific papers. *Current Contents*, 1, 5-7.

Martín-Martín, A.; Ayllón, Juan M.; Orduna-Malea, E. & Delgado López-Cózar, E. (2014). Google Scholar Metrics 2014: a low cost bibliometric tool. *EC3 Working Papers*, 17.

Martín-Martín, A., Ayllón, Juan M., Delgado López-Cózar, E., & Orduña-Malea, E. (2015). Nature 's top 100 Re-revisited. *Journal of the Association for Information Science and Technology*, 66(12), 2714–2714.

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.

Orduna-Malea, E., & Delgado López-Cózar, E. Dimensions: re-discovering the ecosystem of scientific information. *El Profesional de la Información*, 27(2), 420-431.

Orduna-Malea, E., Martín-Martín, A., Ayllón, Juan M., & Delgado López-Cózar, E. (2016). *La revolución Google Scholar : Destapando la caja de Pandora académica*. Granada: Universidad de Granada.