How many academic documents are visible and freely available on the Web?

Enrique Orduña-Malea[1], Juan Manuel Ayllón[2], Alberto Martín-Martín[2], Emilio Delgado López-Cózar[2]

[1] EC3: Evaluación de la Ciencia y de la Comunicación Científica, Universidad Politécnica de Valencia (Spain)
[2] EC3: Evaluación de la Ciencia y de la Comunicación Científica, Universidad de Granada (Spain)

A DIGEST OF

Khabsa, M. & Giles, C. Lee (2014). "The number of scholarly documents on the public web". Plos One, v. 9, n. 5, e93949. doi:10.1371/journal.pone.0093949 http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0093949

SUMMARY

How many academic documents are visible and freely available on the Web?
This is nothing more and nothing less than the tricky question that Khabsa and Giles intends to answer in their work recently published in Plos One in May 2014, with which the review series of Google Scholar Digest starts.

The inclusion of this work in our newsletter is justified not only because Google Scholar is used as a source, but also because it tries to partly estimate its size, though limited to documents written in English.

The structure of this digest consists of the research questions raised, the design and methodology, and the principal results obtained. Finally a brief discussion is offered as well, providing some open questions that this work suggests, related in particular to the characteristics of Google Scholar as a source, and in general to its use in the study of science and scientific communication.

KEYWORDS

Google Scholar / Microsoft Academic Search / Academic web / Open Access

Grupo de Investigación EC3
Evaluación de la Ciencia y de la
Comunicación Científica

EC3's Document Serie:
EC3 Google Scholar Digest Reviews Nº 1

Document History
Version 1.0, Published on 27 May 2014, Granada

Cited as
Orduña-Malea, E.; Ayllón, J.M.; Martín-Martín, A.; Delgado López-Cózar, E. (2014). "How many academic documents are visible and freely available on the Web?". EC3 Google Scholar Digest Reviews, n. 1.

Corresponding authors
Emilio Delgado López-Cózar: edelgado@ugr.es; Enrique Orduña-Malea: enorma@upv.es

1. DIGEST

RESEARCH QUESTIONS

- Can we estimate how many academic papers are circulating on the web?
- Can we estimate how many of them are freely available?
- Are there differences between scientific fields and disciplines?
- How many documents written in English does Google Scholar cover?

METHODOLOGY

Unit of analysis	The calculations refer only to academic papers, which are defined by the authors as: journal and conference papers, dissertations and masters theses, books, technical reports and working papers. Patents are excluded.
	The estimate is limited to documents written in English.
	Inferences are based on studying the coverage of two major academic search engines: Google Scholar and Microsoft Academic Search.

Sample 1: estimating the number of scholarly documents

- 150 English written documents from MAS; 10 of the most cited documents in each of the fifteen fields are randomly sampled, considering only documents with less than 1,000 citations.
- The 15 disciplines used are: Agriculture Science, Arts and Humanities, Biology, Chemistry, Computer Science, Economics and Business, Engineering, Environmental Sciences, Geosciences, Material Science, Mathematics, Medicine, Physics, Social Sciences, and Multidisciplinary.
- Incoming citations to the 150 selected documents: 41,778 citations were obtained from MAS and 86,870 citations from Google Scholar.

Design 1

The number of scholarly documents available on the web is estimated using the Lincoln-Petersen method (capture/recapture): $\frac{R}{M} = \frac{C}{N}$

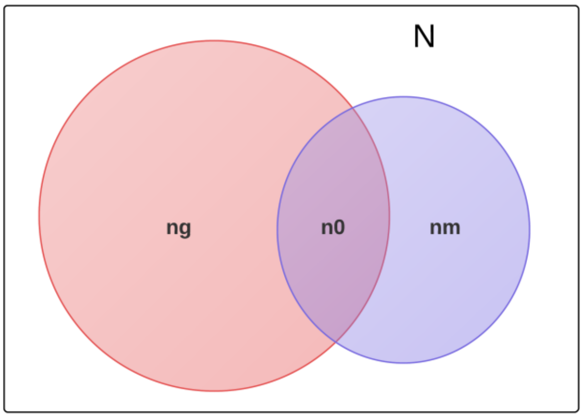

Figure 1. Overlap between GS and MAS used to Capture/Recapture method
Data source: Khabsa & Giles (2014)

Analogies used:	N = size of GS + size of MAS
	M (elements captured in first sample): size of GS
	C elements captured in second sample): size of MAS
	R (elements recaptured in second sample): overlap between GS and MAS

To compute the overlap between Google Scholar and MAS (in general and per discipline) the Jaccard similarity index was used.

To confirm the reliability of this method, the Poisson regression method based on capture/recapture was also used, as explained in the appendix of the article, which is available at:
http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0093949#pone.0093949.s001

Sample 2: estimating the number of free available scholarly documents

1,500 documents from MAS; 100 documents belonging to each field, with at least 1 citation.

Design 2

Whenever a direct link to the full text of a particular search result is available, Google Scholar displays this link next to the search result.

Each of the 100 documents per discipline collected from MAS is searched on Google Scholar, computing if the direct link to the document already exists.

The estimate percentage of freely available documents per discipline is multiplied by the estimated size of the field (obtained in design 1) to obtain the total number of free available documents per field.

Period: All

Data collection date: January 10–12, 2013

RESULTS

1. The number of scholarly documents, published in English, available on the web is roughly 114 million (Figure 2).
2. Google Scholar is estimated to contain 99.3 million documents, which is approximately 87% of the total number of scholar documents on the web (Figure 2).
3. Google Scholar is more than twice as large as the nearest alternative, since MAS and Web of Science are both reported to have fewer than 50 million records (Figure **2**).

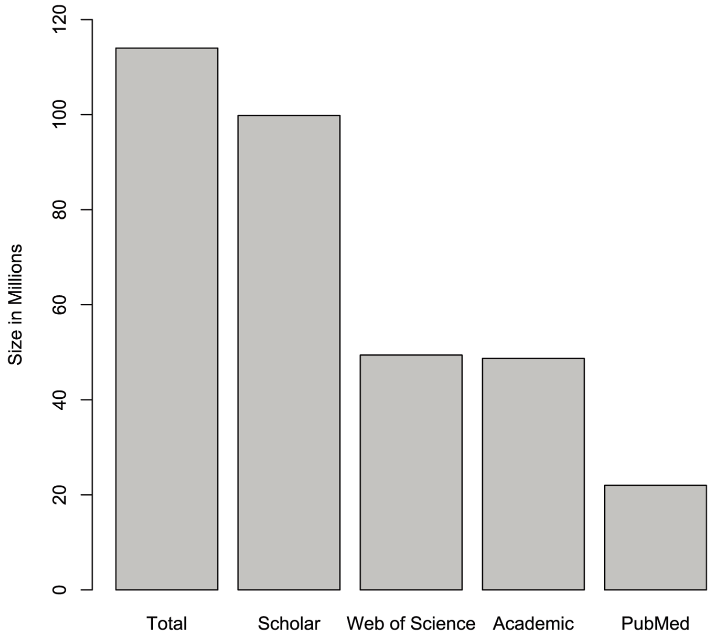

Figure 2. Size of different academic search engines and databases (English documents).
Data source: Khabsa & Giles (2014)

4. The superiority of Google Scholar especially on Multidisciplinary, Social Sciences, Arts & Humanities, and Physics (Figure **3**).

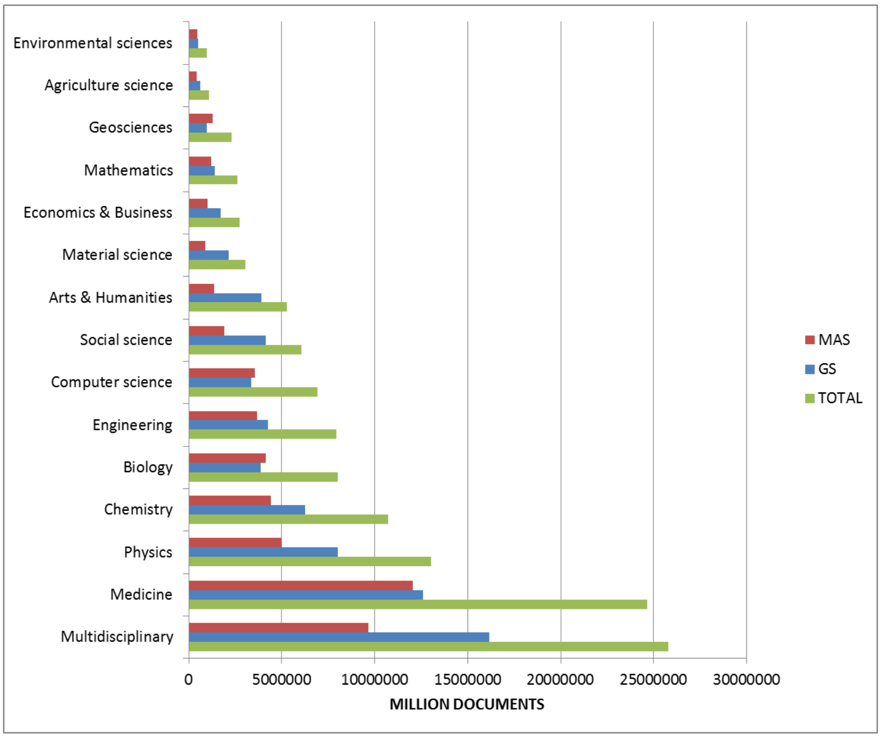

Figure 3. Size of different academic search engines and databases (Lincolm-Petersen)
Data source: re-elaborated from Khabsa & Giles (2014)

5. Approximately 27 million documents (24%) are freely available since they do not require a subscription or payments of any kind in order to access them. 1 in 4 of the web accessible scholarly documents are freely and publicly available (Figure 4).
6. Estimates of open access documents differ significantly for specific academic fields, to the point that some fields have a 4 times greater percentage of freely available documents than others (Figure 4).

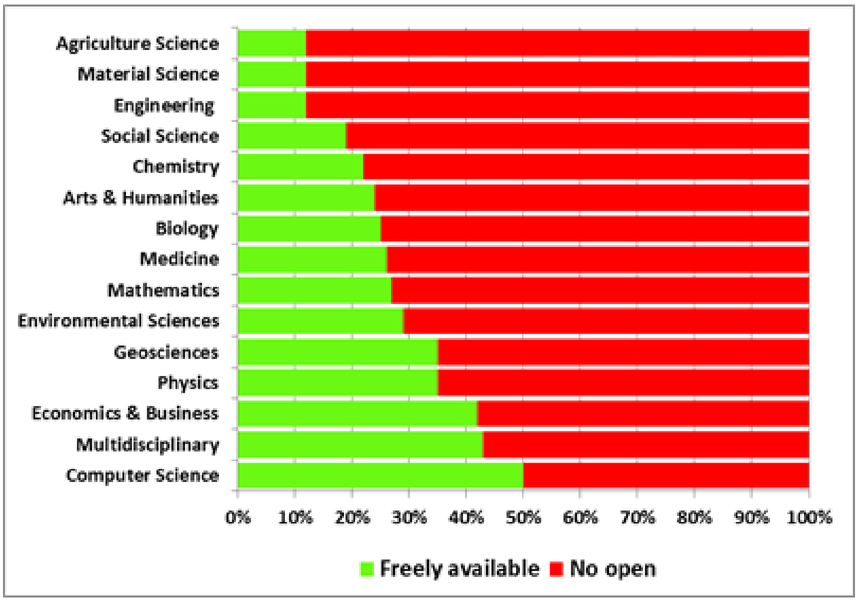

Figure 4. Percentage of publicly available documents according to scientific fields
Data source: re-elaborated from Khabsa & Giles (2014)

2. DISCUSSION

In the light of the results obtained in this seminal article, the following open questions arise:

Is Lincolm-Petersen method a valid procedure to obtain the total population of the public academic web?

The problem is not only related to the possible growth of the population among samples, or the condition on the equal probability of each element to be recaptured, but also to the assumption that each sample is applied to different universes (Google Scholar and Microsoft Academic Search).

The authors use a complementary method obtaining similar results, and this reinforces the results. In any case, a reasonable uncertainty exists.

Can we measure the size of the academic web through Google Scholar and Microsoft Academic Search?

By using the Lincolm-Petersen method, another assumption is made: N (i.e., the scholarly academic public web) is considered to be the summation of Google Scholar and Microsoft Academic Search. On one hand this issue keeps out other databases, such as Google Books (it is well-known that Google Scholar and Google Books databases do not match exactly) or CiteseerX, among others, although we are aware that this missing results are probably low and statistically insignificant. On the other hand there is a more fundamental concern: the low indexation of institutional repositories on Google Scholar (an issue empirically tested, to be published soon), which may indeed influence in the calculated percentages of publicly available documents on Google Scholar.

Moreover, the research assumes as valid the universe of Microsoft Academic Search (48.7 million of documents, as for January 2013). Notwithstanding, the information about the total size of MAS is confusing at present. Microsoft Azure Marketplace shows (as for May 2014) 39.85 million documents, which do not match with the data used in Khabsa & Giles research; from the information gathered on the Web, we can estimate 45.3 million documents, and 45.9 million documents if a query is performed manually in the website platform (as for May 2014). How can this disperse information affect the calculation of “N”?

Can we measure the size of the academic web through search engines?

The authors echo themselves about this limitation: search engines impose a restriction on the number of retrievable results for all type of queries, unless an Application Programmable Interface (API) is provided (and Google Scholar does not provide an API at the moment).

Accessing to only the first 1000 documents may bias the sample in an unknown way (and maybe differently in each discipline), although we can assume (though not demonstrate) that these records contain the more formalized, visible, and more circulating and cited documents. Moreover, this statistical error is equally distributed to all 15 samples, reducing its effect.

Can we measure the size of the academic web through citing documents?

The research design performed by the authors is novel and brilliant. It is based on the gathering of citing documents to a sample of 150 articles.

This procedure has several advantages, among others:

a) The search engine is forced to query its entire database to find all documents that match with a citation to any of the documents of the sample.
b) The citing documents are diverse in typology as Google Scholar gathers many types of scholarly material

Nonetheless, this introduces some disadvantages as well:

a) Citing documents can be in all languages. The authors identify that 98% of citing papers in MAS are written in English (a correction is applied instead of eliminating non-English documents), but this percentage in Google Scholar must be lower, and it is not indicated in the estimation of the size of Google Scholar. Probably all documents written in languages other than English should have been avoided if the estimation of the English academic web was the objective. For this reason, we believe that the calculation is not limited to the English academic web space.
b) The sample is composed by articles, whereas citing documents are diverse in typology. If the target of this research had been calculating the number of "articles" included on Google Scholar, then this procedure would have been appropriate (after the elimination of citing documents other than articles). If the purpose were to calculate the size of the entire database, a sample composed uniquely by articles is not representative of Google Scholar, and this is the limitation of using Lincolm-Petersen with two different universes.
c) Despite the fact that the sizes of each discipline are quite different, as showed in the results (Figure 3), the sample is uniform for each field (10 articles). This can introduce an important bias in the estimation.

All these questions lead us to consider that the estimation of Khabsa & Giles is probably lower than the real size of the academic web. In any case we consider it a seminal research due both to its novelty in the research design and the implications to science and scientific communication research.

For this reason, EC3 Research Group is working to offer the following working papers, which will be released soon:

- About the size of the public academic web and Google Scholar
- About the size of the Open Access literature.