

4. Google Scholar as a Data Source for Research Assessment

Emilio Delgado López-Cózar , Enrique Orduña-Malea , Alberto Martín-Martín 

The launch of Google Scholar (GS) marked the beginning of a revolution in the scientific information market. This search engine, unlike traditional databases, automatically indexes information from the academic web. Its ease of use, together with its wide coverage and fast indexing speed, have made it the first tool most scientists currently turn to when they need to carry out a literature search. Additionally, the fact that its search results were accompanied from the beginning by citation counts, as well as the later development of secondary products that leverage this citation data (such as Google Scholar Metrics and Google Scholar Citations), made many scientists wonder about its potential as a source of data for bibliometric analyses. The goal of this chapter is to lay the foundations for the use of GS as a supplementary source (and in some disciplines, arguably the best alternative) for scientific evaluation. First, we present a general overview of how GS works. Second, we present empirical evidences about its main characteristics (size, coverage, and growth rate). Third, we carry out a systematic analysis of the main limitations this search engine presents as a tool for the evaluation of scientific performance. Lastly, we discuss the main differences between GS and other more traditional bibliographic data-

4.1	The Origins of Google Scholar	95
4.2	Basic Functioning of Google Scholar	97
4.2.1	The Academic Search Engine	97
4.2.2	What Sources Does Google Scholar Index?	98
4.2.3	Google Scholar's Official Bibliometric Products	100
4.3	Radiographing a Big Data Bibliographic Source	102
4.3.1	Size	102
4.3.2	Coverage	105
4.3.3	Growth Rate	117
4.4	Google Scholar's Data for Scientometric Analyses	119
4.4.1	Errors in Google Scholar	119
4.4.2	Google Scholar Limitations	119
4.5	The Expanded Academic World of Google Scholar	121
4.6	Final Remarks	123
	References	125

bases in light of the correlations found between their citation data. We conclude that GS presents a broader view of the academic world because it has brought to light a great amount of sources that were not previously visible.

4.1 The Origins of Google Scholar

The development of the field of bibliometrics has always been reliant on the availability of large-scale sources of metadata about scientific publications, which are ultimately the raw materials used by bibliometricians to carry out their analyses [4.1]. The creation of the first citation indexes by Eugene Garfield (Science Citation Index in 1964, Social Science Citation Index

in 1973, and the Arts and Humanities Citation Index in 1978) turned out to be a crucial turning point that enabled the development of modern bibliometric studies. His novel approach to bibliographic information systems opened the way to a completely new way of assessing scientific performance [4.2].

By indexing not only the articles published in scientific journals, but also the bibliographic references included in these articles, it was possible for the first time to track the relationships between scientists, journals, and institutions through the main tangible outputs these entities produce: scientific documents. Thus, the

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-02511-3_4) contains supplementary material, which is available to authorized users.

databases of the Institute for Scientific Information (now part of Clarivate's Web of Science; WoS) became the first, and for a long time the only available sources of data for bibliometric analyses, exercising an almost absolute monopoly in this field. The use of other specialized databases (such as MEDLINE, Chemical Abstracts, Inspec, or Biosis) for bibliometric purposes was testimonial, since they did not offer citation data nor other fields (e.g., full affiliations of the authors), which are vital to produce useful bibliometric reports.

It was not until the first decade of the twenty-first century that this monopoly was seriously challenged. Elsevier launched its Scopus citation database on November 3rd, 2004. Just a few weeks later (November 18th) GS was also launched. Scopus was conceived as a traditional subscription-based bibliographic database (which indexed a specific set of journals and conference proceedings) and was clearly a direct competitor of WoS. GS departed entirely from this approach, following instead the path of its big brother, the Google search engine, a decision that greatly impacted its design and coverage.

Simply put, GS is a specialized search engine that only indexes academic documents [4.3, 4]. GS's spiders constantly crawl the websites of universities, scientific publishers, topic and institutional repositories, databases, aggregators, library catalogues, and any other web spaces where they might find academic-like materials, regardless of their subject or language. GS indexes documents from the whole range of academic document types (books, book chapters, journal and conference articles, teaching materials, theses, posters, presentations, reports, patents, etc.). Unlike the cumulative and selective nature of WoS and Scopus, GS is dynamic; it reflects the state of the web as it is visible to its search robots and to the majority of users at a specific moment in time. Documents that for any reason become unavailable on the web will eventually disappear from GS too, as will the citations they provided to other documents [4.5].

GS, like the Google search engine before it, achieved instant success among users worldwide. The reason is easy to understand: GS finds most of the scientific information that circulates around the web in an easy and fast manner. Perhaps most importantly, it is free, unlike most of the bibliographic databases that existed before it, which are often only accessible through costly national or university-level subscriptions.

GS is currently the tool most users first turn to when they need to carry out a literature search. This has been evidenced by numerous studies [4.6–11]. *Bosman and Kramer's* study [4.7] is the most recent and large-scale study on the matter. They conducted a survey on the changing landscape of scholarly communication be-

tween May 2015 and February 2016, obtaining more than 20 000 responses from researchers, university students, librarians, and other members of the scholarly community. With respect to the question *What tools do you use to search literature?* GS emerged as the preferred option, selected by 89% of the respondents, followed at a great distance by WoS (41%), Pubmed (40%), others (36%), and Scopus (26%).

Since its launch in 2004, GS's interface has gone through several renovations, but the really important changes (updates to its algorithms, coverage) usually happened under the hood, unbeknownst to most users. Some developments, however, did not go unnoticed. We are referring, of course, to the creation of its two secondary products: Google Scholar Citations (GSC) and Google Scholar Metrics (GSM). GSC was launched in July 2011 and provided a platform in which users could easily create an academic profile by pulling their publications from the data available in GS. Most interestingly, these profiles also displayed several author-level bibliometric indicators [4.12]. GSM was born on April 2012 as a ranking of scholarly publications according to their h-index calculated from GS data. This tool provides an easy way to identify the most influential publications (journals, proceedings, and repositories) and articles published in recent years [4.13].

Although these two tools never lose sight of GS's main purpose (they are intended to serve as search tools, one to find relevant researchers, the other to find influential articles and publications), they use bibliometric indicators as an evidence of relevance. For the first time, the GS team decided to put the citation data available in GS to other uses. Until the creation of those products, citation counts were only used as one of the parameters to rank documents in a search and a search aid for users (*Cited by* links in GS).

The availability of citation data in GS and its secondary products GSC and GSM has attracted the attention of some bibliometricians, and even scientists from other fields, who have realized that the data available in GS provides a much more comprehensive insight into the impact publications have on their respective academic communities than the data available in other citation databases. However, the use of GS for bibliometric purposes was never one of the applications GS's developers intended for this product, and so an exhaustive critical evaluation that analyzes its suitability for bibliometric analyses is necessary.

In order to do this, this chapter first presents a general overview of how GS and its secondary products GSC and GSM work, their inclusion policies, and how they respond (results offered) when specific stimuli (user queries) are applied. Second, we present empirical evidences regarding its size, evolution (growth rate, indexing speed), coverage (publishers, repository

ries, bibliographic databases, catalogues), and diversity (subjects, languages, document types). Third, we carry out a systematic analysis of the main limitations this search engine presents as a tool for the evaluation of scientific performance. Lastly, we discuss the main differences between GS and other traditional bibliographic databases in light of the correlations found between

their citation data at the level of authors, documents, and journals.

The online version of this chapter is complemented by electronic supplementary material (ESM) in which the sources of the empirical analyses are made available, and further documents illustrating the findings on GS are provided [4.14].

4.2 Basic Functioning of Google Scholar

In this section, we first present a concise but accurate description of how the GS search engine works. Secondly, we describe its main inclusion criteria (both for sources and, especially, for documents). Lastly, we will briefly outline GSC and GSM.

4.2.1 The Academic Search Engine

Classic bibliographic databases usually work on the principle of whitelists. They first generate a whitelist of sources that meet some specific criteria (quality, subject scope, ...) and then index all the publications that appear in these sources. The historical tendency to select some specific sources (mainly journals) and not other channels for the dissemination of academic results (conference proceedings, books, reports, etc.) responds mainly for two reasons. First, it is a question of efficiency, usually referred to as Bradford's law of scattering [4.15], thanks to which we know that for any given topic, a small core of journals provides most of the articles on that topic. When faced with technological and economic constraints, maximizing returns by selecting only the core of journals that will be most useful for a given purpose seems a logical and natural response. The other reason has to do with the evaluative use of these databases. Due to their visibility and prestige, most authors want to publish their articles in these core journals, increasing the competition to get a manuscript accepted in these journals. The limited space for publication of the printing era, as well as the higher standards to which articles are held in these journals are what help project their image of prestige. To publish an article in a core journal is a difficult task, something that only the best researchers manage to do. In the same way, receiving a citation from an article published in a core journal also lends prestige to the cited article and its authors. This is the road to research excellence.

It goes without saying that this traditional approach (which prioritizes the optimization of resources and excellence) is not without its merits and has played an important role up until now. However, the irruption of

GS represents a break from this paradigm. Unlike traditional bibliographic databases, which are selective by nature, GS parses the entire academic web, indexing every scholarly document it finds, regardless of its quality, and it does not differentiate between peer-reviewed and non-peer-reviewed content. GS is, rather, an academic search engine [4.3] with a bibliographic database that grows in a (mostly) unsupervised manner, and which has one clear purpose: facilitating the discovery of academic literature for everyone worldwide.

This unsupervised indexing process is possible thanks to the automated bots (sometimes also called spiders) that GS deploys throughout the web, similarly to how the Google search engine crawls the web as well. In GS, these bots are trained to locate academic resources, index their full texts (whenever possible), and extract their bibliographic descriptions (metadata). The process ends with the automated creation of a bibliographic record that is ready to be included in a search engine results page (SERP) when it is deemed relevant for a particular query. In order for a particular academic website to be successfully indexed in GS, certain technical requirements must be met, i. e., bots must be allowed to enter the website, there must be an easy-to-follow route to the article pages, and certain metadata must be available on these article pages. More detailed information can be found on the GS help pages [4.16].

When GS's bots are able to access the full text of the documents (either because the resource is openly available on the web, or thanks to the special agreements GS has with most commercial publishers), they also extract the list of cited references from each document. Thus, they are able to link citing and cited documents, which is how they can calculate citation counts.

When a list of cited references is parsed and processed, GS tries to find matches to those documents in its database. If it finds a match, it links the citing and cited document, and the cited document will have one more citation. However, if it does not find a match for a particular cited reference, the system will create a new bibliographic record of the type [CITATION], to which the citing document will be linked. These

records are also displayed in SERPs, although it is possible to exclude them, as their bibliographic information is often incomplete, and users will not be able to access their full text. A *[CITATION]* record can become a full-fledged record if GS finds another version of the document on the web (because someone deposits it on a repository, or it becomes available from a publisher, etc.), and merges the two versions. Needless to say, this entire process is also completely automated.

Lastly, GS considers a wide range of parameters for ranking documents in the SERPs, such as

weighing the full text of each document, where it was published, who it was written by, as well as how often and how recently it has been cited in other scholarly literature. [4.17]

However, the detailed set of parameters and the weight each of them has in the ranking algorithm is not publicly available.

4.2.2 What Sources Does Google Scholar Index?

The previous section makes clear the distinction between GS and traditional bibliographic databases. In their own words, “we index papers, not journals” [4.18]. However, this statement is only partially true. We would rather define GS as a database that indexes web sources. Moreover, it deliberately includes some document collections (e. g., patents, court opinions, and *[CITATION]* type records).

GS crawls a wide variety of web sources and indexes everything from those sources that it identifies as academic documents. That is why GS includes all documents regardless of subject, language, country, or year of publication, and document type. The procedure GS follows to index new documents can be summarized in three steps, which are described below.

Step 1: Compilation of Sources

Over the years, GS has compiled a huge list of sources, ranging from websites of academic institutions (higher education institutions, national research councils, commercial publishers, private companies, professional societies, non-governmental organizations, etc.), to other discovery tools (bibliographic databases, catalogues, directories, repositories, other search engines) available across the web. These sources, which are the most likely to host academic content, are the ones that shape the academic web. Once they add a source to their private master list, GS’s spiders will visit it periodically to check whether new documents have been added and also to verify that the documents indexed in the past are still available.

Besides the already mentioned academic sources, anyone can request that their website be considered for inclusion in GS. They are prepared to index websites that run in most of the common repository platforms (DSpace, EPrints, ...), journal platforms (OJS—open journal systems), and also simple personal websites.

Since GS’s main objective is to facilitate content discovery, the sources must not require users to install additional applications, to log in, use Flash, JavaScript, form-based navigation, or any other kind of unreasonable methods to access the documents. In addition to that, the website should not display popups, interstitial ads, or disclaimers. They specifically state that

all those websites that show log-in pages, error pages, or bare bibliographic data without abstracts will not be considered for inclusion and may be removed from GS.

Step 2: Document Types

The next step is to index the academic documents available in each source. GS does not index all the documents in a source, only those that are academic in nature.

GS states that they cover mostly “scholarly articles, journal papers, conference papers, technical reports, or their drafts, dissertations, pre-prints, post-prints, or abstracts”. However, content such as “news or magazine articles, book reviews, and editorials” is not considered appropriate for GS. However, appropriateness is not really a constraint, and the documentation also states that “shorter articles, such as book reviews, news sections, editorials, announcements and letters, may or may not be included”.

These ambiguous declarations are a consequence of the automated way in which the system operates. Let us see a practical example:

The University of Oxford’s official website (<http://ox.ac.uk>) can be considered a reliable source of academic information. GS has added this web domain to its master list of indexable sources (like it does with most universities). However, not all documents hosted in ox.ac.uk are of an academic nature. GS needs to automatically differentiate academic documents from all the rest. To do this, the system applies two approaches:

1. The parser approach: GS uses full-text parsers to identify the structure of documents. Taking advantage of the fact that many academic documents tend to present a fairly standardized structure (title, author names, abstract, body of the article, references, ...), detecting whether or not a document is academic is often possible, although errors do occur.

2. The location approach: GS automatically indexes all the documents hosted in specific locations where it is reasonable to expect that all documents will have an academic nature, i. e., institutional repositories.

For this reason, despite what is stated in GS's documentation, it is possible to find a great range of document types in GS. Documents are usually stored in the HTML or PDF format. Parsing the structure from these documents is not enough to detect their specific typology (article, book chapter, conference paper, etc.) when additional metadata is not available. Moreover, once it has been decided that all content from a given source will be indexed, the actual document type stops mattering. For example, in GS we can find many book reviews, a document type that is explicitly considered inappropriate (Fig. 4.1).

Step 3: Documents

Lastly, documents themselves must also follow certain guidelines in order to be successfully indexed in GS. Some of them are compulsory, and others are only optional. Failure to comply with these rules may provoke an incorrect indexing of the documents, or, more likely, a complete exclusion from the search engine.

The system requires one URL per document (one intellectual work should not be divided into different files, while one URL should not contain various independent works). Additionally, the size of the files must not ex-

ceed 5 MB. Although documents of a larger size can appear in GS, their full text (including cited references) will be excluded if they do not comply with this rule.

HTML and PDF are the recommended file types. Other document types such as DOC, PS, or PPT are also indexed, but they are a very small minority and they might not be processed as effectively as the others. Additionally, PDFs must follow two important rules. First, all PDF files must have searchable text. If the PDFs are just scanned images, the full texts (including the cited references) will not be processed, since GS's crawlers are unable to parse images. Second, all URLs pointing to PDF files must end with the '.pdf' file extension.

There are also rules regarding the description of the articles through metadata. Some fields are compulsory for all documents (title, authors, and publication date), while others are specific to each document type. GS supports HTML meta tags in various formats: Highwire Press, Eprints, BE Press, PRISM, and Dublin Core (the last one as a last resort, since there are no specific fields for journal title, volume, issue, and page numbers in this format). If no metadata is readily available in the HTML meta tags of the page describing the article, GS will try to extract bibliographic information by parsing the full text of the document directly. For this reason, GS also makes recommendations regarding the layout of the full texts:

- The title, authors, and abstract should all be on the first page of the text file.

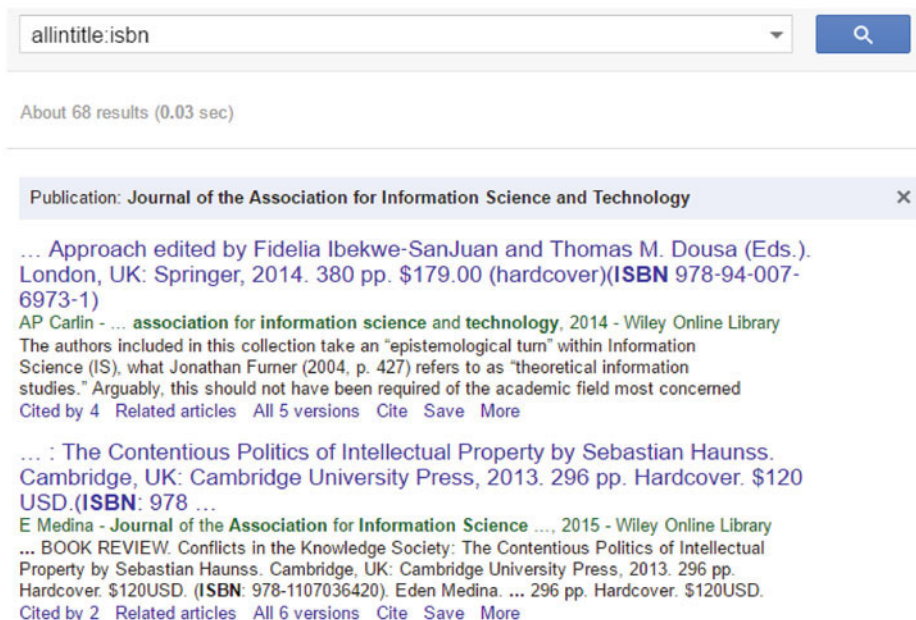


Fig. 4.1 Book reviews published in the *Journal of the Association for Information Science and Technology* (JASIST) and indexed in GS

- The title should be the first content in the document and no other text should be displayed with a larger font size.
- The list of authors should be listed below the title, in smaller font size, but larger than the font size used for the normal text.
- At the end of the document, there should be a separate section called *References* or *Bibliography*, containing a list of numbered references.

Lastly, the abstract of the document should be visible to all users that visit the article page (regardless of whether or not they have access to the full text of the document) without needing to click on any additional links or to log in. If this requisite is not met, it is likely that the document will not be indexed in GS.

4.2.3 Google Scholar's Official Bibliometric Products

GS has developed two secondary products that make use of both the bibliographic and citation data available in its core database. The first one (GSC) focuses on researchers, while the second one (GSM) focuses on journals and articles. This section describes each product briefly and discusses the bibliometric indicators they provide.

Google Scholar Citations

GSC was officially launched on November 16th, 2011. This tool is an academic profile service meant to help researchers maintain an up-to-date list of their publications without much effort (it is updated automatically), and it also facilitates searches of people (rather than documents) who are experts in any given academic topic.

First, GSC profiles contain structured personal information (name of the researcher, affiliation, and research interests). Second, the profiles show a list of all the publications written by the researcher. For each of the publications, both bibliographic (authors, title, source, year of publication) and citation data (number of citations and link to the list of citing documents) are offered. By default, documents are sorted decreasingly by number of citations, although they can also be sorted

by year of publication or by title. Third, the profile also provides several author-level indicators (Table 4.1). These indicators are calculated considering two different time frames: first, without any time restriction (which is useful for comparisons of senior scholars), and second, considering only citations received in the last 5 years (which is useful for comparisons of early-career researchers). It is important to keep in mind that these indicators are calculated automatically from the data in the publication list, without any sort of human supervision.

In GSC, users have access to a document search tool that enables them to find their publications by means of author name searches (it is possible to search as many name variants as necessary), or by known document searches (usually title searches). After all documents have been found, researches may merge versions of the same document that GS has not been able to detect and fix bibliographic errors manually. All these operations only affect the researcher's profile and not other co-authors' profiles or the results in the GS search.

The platform also offers additional services, such as personalized alerts, lists of co-authors, areas of interests, list of authors by institution, etc. Therefore, authors can track the impact of their papers and other researchers' papers according to the data available in GS and be instantly informed of new papers published by other authors. All these features make GSC a powerful and free research monitoring system.

This product may be viewed as a first step in the transition from an uncontrolled database to a better-structured system where authors, journals, institutions, and areas of interest go through human filters. Nevertheless, the platform still lacks some essential features, like the identification of document types. This information can be defined by the owner of the profile on the document edit page, but it is not visible to other users visiting the profile. Another important issue is that author affiliations are not available at the level of documents. Although the affiliation field of the profile can be modified, it is only possible to display one affiliation at a time. However, researchers may change affiliations, and it would not be fair to ascribe all documents by a researcher to only one institution, if some of them were published while that researcher was working at other

Table 4.1 GSC's author-level metrics

Metrics	All	Last 5 years
Citations	Number of cites an author has received	Number of cites an author has received in the last 5 complete calendar years
h-Index	The largest number h such that h publications have at least h citations	The largest number h such that h publications have at least h new citations in the last 5 years
i10-Index	Number of publications with at least 10 citations	Number of publications with at least 10 citations in the last 5 years

institutions. Limitations like these diminish the usefulness of the platform for bibliometric studies (for more limitations, please see ESM, Appendix VI).

Google Scholar Metrics

GSM was launched on April 1st, 2012 and can be defined as a hybrid between a bibliographic and bibliometric product that presents a ranking of journals according to bibliometric indicators calculated using citation data from recently published articles in those journals. If GS represented a paradigm shift in the market of bibliographic databases, GSM accomplished something similar with respect to journal rankings, especially when compared to products like journal citation reports (JCR), the SCImago journal and country rank (SJR), or journal metrics [4.19]. GSM is an original product for various reasons:

- *Inclusion policies:* GSM only covers journals that have published at least 100 articles in the last 5 years and which have received at least one citation for any of those articles.
- *Coverage:* Apart from journals, GSM also covers some conference proceedings from computer science and electrical engineering, and preprint repositories. Other typologies like court opinions, books, and dissertations are explicitly excluded.
- *Sorting criteria:* Sources are sorted by their h5-index (h-index for articles published in a given 5-year period). The use of an h-index variant instead of a formula similar to the journal impact factor (JCR), SJR, SNIP (source normalized impact per paper), or CiteScore (journal metrics) is probably one of the most distinct features of this product. A description of the indicators available in GSM is presented in Table 4.2. For each journal, only the articles that contribute to the h5-index are displayed (h5-core). These articles are also accompanied by their citation counts, and the list of citing documents to each article is also available.
- *Categorization of sources:* The first variable of categorization is the language of publication. In the version available at the time of writing (launched in summer 2016, covering the period 2011–2015), the following languages were covered: English, Chinese, Portuguese, Spanish, German, Russian, French, Japanese, Korean, Polish, Ukrainian, and Indonesian. For each of these languages, except for English, the ranking displays the top 100 sources according to their h5-index. For English sources, a subject classification is also provided (Table 4.3).

Table 4.2 Google Scholar Metrics

Metrics	All
h5-Index	The largest number h such that at least h articles in that publication were cited at least h times each in the last 5 years
h5-Core	The set of articles from a journal with a citation count above the h5-index threshold
h5-Median	Median of the distribution of citations to the articles in the h5-core.

Table 4.3 Categories and number of subcategories in GSM

Categories	Number of subcategories
Business, economics, and management	16
Chemical and materials science	18
Engineering and computer science	58
Health and medical science	69
Humanities, literature, and arts	26
Life sciences and earth sciences	39
Physics and mathematics	24
Social sciences	52

The classification scheme is made of 8 main categories and 302 subcategories. In each of the categories and subcategories, the number of results displayed is limited to the top 20 journals according to their h5-index.

It should be pointed out that some subcategories are included in several categories (library and information science, for example, is included both in engineering and computer science, and in social sciences), and that one source may be included in more than one subcategory.

- *Search tool:* The platform also provides an internal search box, which enables users to locate journals that are not included in any of the general rankings. Users can carry out keyword queries, which will return sources with names that match the query. Each response to a query contains a maximum of 20 sources, also sorted according to their h5-index.

The peculiar features of this journal ranking have been tested, and numerous limitations found, such as a lack of name standardization, irreproducible data, and a questionable mix of publication typologies [4.20]. Nevertheless, the product has improved since its first editions, revealing itself as a potential source for the evaluation of journals in the areas of humanities and social sciences [4.21, 22].

4.3 Radiographing a *Big Data* Bibliographic Source

The goal of this section is to provide empirical data about the bibliographic properties of GS as a database. Three aspects will be discussed: size, coverage, and growth rate.

4.3.1 Size

One of the most crucial aspects that make us consider GS a *big data* source is the issue of determining its size. Unlike Scopus or WoS, highly controlled databases where finding out the total number of records only requires a simple query, GS is a search engine that presents what is available in the academic web at a specific moment in time. However, the web is not only dynamic, but also unstable and uncontrollable.

Therefore, the methodological difficulties of ascertaining the size of search engines are related to stability problems [4.23, 24], precision of the results obtained [4.25, 26], and the degree of permanence and persistence of the resources [4.27–29]. The high instability of search results and the lack of precise search commands have led experts interested in finding the size of a search engine to use methods based on the extrapolation of frequencies of documents available in external sources [4.30].

As regards GS, there are two types of methods to calculate its size: direct methods (based on the execution of queries in the search engine) and indirect methods (estimations based on comparisons with external sources for which more information is known). Among the direct methods, three strategies are worth mentioning:

1. Web domain queries [4.31–33]
2. Year queries [4.3]
3. The so-called *absurd* queries [4.34].

Among the indirect methods, the capture/recapture method [4.35] and the proportion of documents in En-

glish respect to the total [4.34] have been attempted. For these last two methods, additional information must be known about the databases used as a reference. To date, these studies find that direct methods based on web domain queries and absurd queries are the ones that yield higher figures, which are similar in both cases.

Before discussing the calculation of the size of GS, it is appropriate to describe the characteristics of its coverage, since this is the key aspect to understanding the results. GS is currently made of two separate document collections: articles and case laws. The analysis of the latter is outside the scope of this chapter. The article collection is, in turn, divided into source documents and cited references (documents that GS's crawlers have only been able to find as references inside other source documents or certain metadata-only bibliographic databases). Cited references are marked with the text *[CITATION]* in SERPs and can also accrue citations of their own, which are displayed in the same way as for source documents. There is a last document type to which GS gives special attention in its interface: patents.

The integration of source documents and cited references in the same list of results breaks from the way WoS and Scopus handle these types of records, where each collection is displayed separately. In WoS, cited references are accessible from a completely separate search system (cited reference search), while in Scopus, cited references are also displayed separately as *secondary documents*.

There are two types of *[CITATION]* records: *Linked citations* (documents for which only basic bibliographic information—but no access to full-text—has been found in some library catalogue or metadata-only database) and *unlinked citations* (documents that have been cited in source documents and which the system has not been able to find anywhere else on the web).

Table 4.4 shows a compilation of the studies that have provided estimations of the size of GS. As can

Table 4.4 Compilation of studies on the size of GS

Authors	Date	Method	Coverage	Language	Size
Aguillo [4.32]	August 2010	Direct-domains	Articles + citations + patents	All	86 million
Ortega [4.3]	December 2012	Direct-date query	Articles + citations + patents	All	95 million
Khabsa and Giles [4.35]	January 2013	Indirect-cap/recap	N/A	English	99 million
Orduna-Malea et al. [4.34]	May 2014	Indirect-empirical studies	N/A	All	171 million
		Direct-date query	Articles + citations + patents	All	100 million
		Direct-absurd query	Articles + citations + patents	All	170 million
Aguillo [4.33]	January 2017	Direct-domains	Articles + citations + patents	All	194 million
Orduna-Malea et al. [4.36]	March 2017	Direct-absurd query	Articles + citations + patents	All	331 million
		–	Articles	All	184 million
		Direct-domains	Articles	All	197 million

N/A: not available

be expected, the results are affected by the estimation method, date of data collection, languages covered, and specific parameters of the searchers (inclusion or exclusion of cited references and patents).

Aiming to offer results as updated as possible, we replicated the absurd query in March 2017. A series of year queries, combined with the command ‘fsdfs-dgsdh.info’ were carried out, and the number of hits each search yielded was collected. Through this simple procedure, we obtained a total of 184 001 450 source documents. Together with cited references (134 160 570) and patents (13 742 920), bringing the total to 330 804 940 documents.

Figure 4.2 offers a comparison of the size of GS, WoS core collection, and Scopus at two moments in time: 2011 and 2016. According to the most recent data, the coverage of GS seems to be almost three times as large as that of WoScc (2.8 : 1) and Scopus (2.7 : 1).

In order to test the robustness of the results, we compared the results returned by various types of queries

to GS, carried out at different moments in time (Table 4.5). A high correlation was found among the results found for each year for all queries, regardless of whether or not cited references and patents were excluded.

Obviously, all these results are merely approximations of the size of GS. Its exact size cannot be ascertained with precision. Given the magnitude of the numbers the system returns (millions of documents), estimations are the best that we can expect when working with academic search engines.

Although the size of GS is clearly larger than that of other databases, the coverage by years can show us which database has a higher coverage in specific years. Figure 4.3 shows the number of documents by publication year that GS, WoS, and Scopus covered at the time of writing this.

As Fig. 4.3 shows, GS covers many more source documents than the other two databases, both for old material (first half of the nineteenth century) and recent

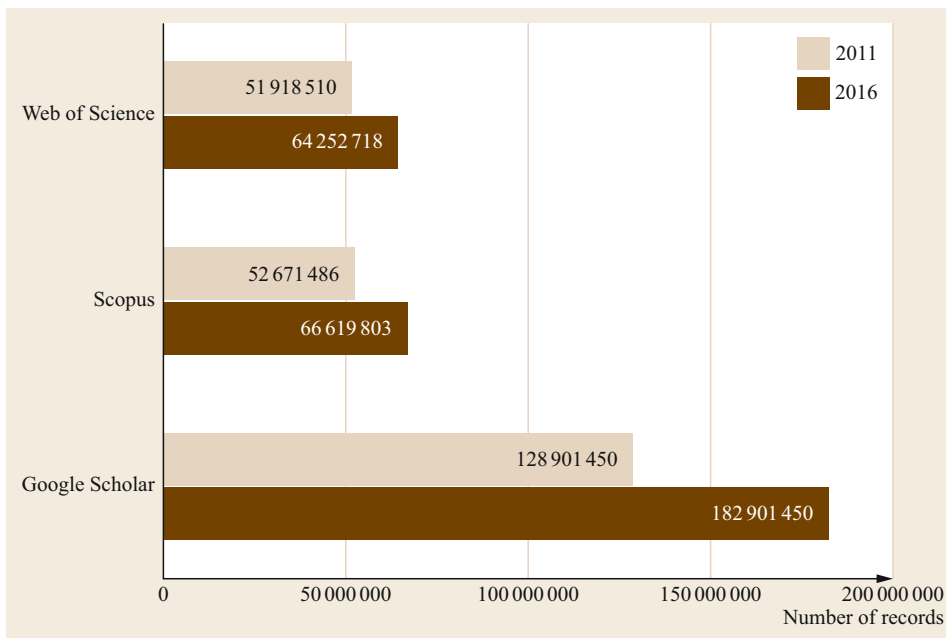


Fig. 4.2
Sectional coverage of GS, WoScc, and Scopus (from origin to 2016, included)

Table 4.5 Correlation among different queries (1800–2013)

Queries	Absurd pure (2017)	Date pure (2017)	Absurd full (2017)	Absurd full (2014)	Date full (2014)
Absurd-pure (2017)	1	0.997	0.992	0.978	0.976
Date-pure (2017)	0.997	1	0.994	0.984	0.983
Absurd-full (2017)	0.992	0.994	1	0.990	0.990
Absurd-full (2014)	0.978	0.984	0.990	1	0.995
Date-full (2014)	0.976	0.983	0.990	0.995	1

Pure: Excluding citations and patents

Full: Including citations and patents

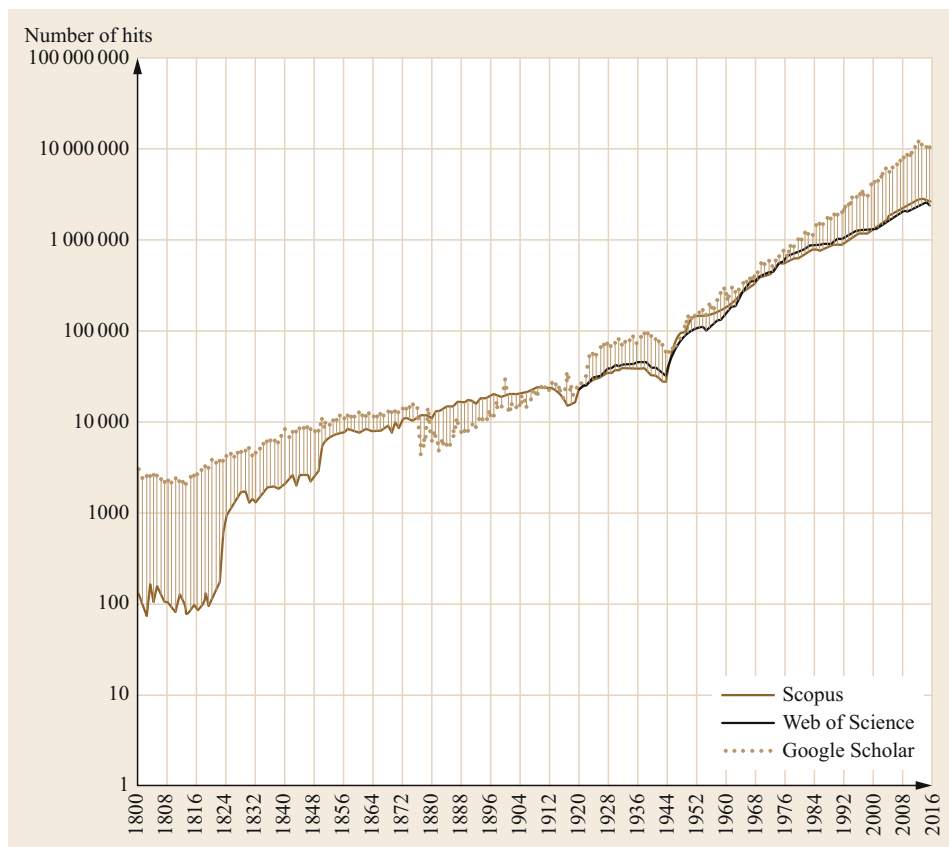


Fig. 4.3 Size of GS, WoScc, and Scopus (1800–2016). Note: The size of GS with respect to citing sources, eliminating cited references, in order to make GS comparable with the remaining databases

material (from the beginning of the twenty-first century onwards). The vast majority of the content covered by these databases was published in the current century (70.4% of all documents).

Up to this point, we have analyzed the size of GS by studying its source documents. However, it is also possible to approach this issue by studying the relationships between documents, that is, the citations themselves. As was previously said, an indirect method to find out the size of GS is to use other databases as a benchmark. Differences in citation counts in documents that are covered both by GS and the database used as a reference can be considered an indication of the underlying differences of their document bases: If a database has been able to find more citations than another database for a particular set of documents, its document base must be bigger (assuming the citation tracking mechanisms of the databases under comparison have roughly the same effectiveness).

Numerous studies have been published on this issue, most of them comparing GS to WoS and/or Scopus. The units of measurement used in these studies are varied:

1. Number of publications
2. Number of citations
3. h-indexes
4. Percentage of unique citations in each database.

A simple summary of these empirical results can reveal to which degree the size of GS's document base is bigger than that of the traditional databases, and how these differences have changed over time.

A table with a list of 63 empirical studies that address the issue of the size of GS as compared to WoS and/or Scopus is available in the electronic supplementary material [4.14, see Appendixes I, III, and IV]. For each study, the sample, discipline studied, unit of measurement used, results obtained, and ratio between the results for GS and the other databases are given. The results must be interpreted with caution, because the ratios depend to a large degree on the disciplines studied, the geographic scope (international or national), and the linguistic scope. Moreover, they can also be affected by the size of the samples (which, in general, tended to be very small).

From this meta-analysis, we can confirm several facts:

- To date, there are few studies that compare GS, WoS, and Scopus at the same time. The most frequent comparison is between GS and WoS.
- The vast majority of studies make comparisons on the basis of citations (54), way ahead of those that use documents—usually articles—(24), or indicators like the h-index (12).
- Out of the 63 studies, only 8 yielded results where WoS and/or Scopus surpassed GS in terms of size. All these studies analyzed STEM (science, technology, engineering and mathematics) fields, like chemistry. The majority of studies show that, for any given set of documents covered both by GS and WoS/Scopus, the first is able to find a higher number of citations.
- The greatest differences in favor of GS are found in the humanities and the social sciences. As regards STEM disciplines, GS is still able to find more citations, but the differences are less marked.
- The differences are greater when comparisons are made on the basis documents written in languages other than English.
- The differences between GS and the other databases seem to increase in the more recent studies, which might indicate an even broader coverage in GS respect to the other databases in recent years.
- Ratios of GS indicators to indicators from other databases are lower when comparisons are made on the basis of the h-index or the number of publications, and they are higher when citations are used instead. Ratios are even more in favor of GS when only unique citations are analyzed (citations found in one database and not in the others). These unique citations in GS, that is, its ability to find citations that no other database is able to find (not limiting itself to strictly scientific sources, but covering all the academic and professional sources it can find), are what make GS truly unique.

So as to provide a broader and more updated perspective on this issue, we have analyzed two large samples of documents covered both by GS and WoS. The results are consistent with the previous studies.

According to the samples analyzed in Table 4.6, the ratio of GS citations to WoS citations ranges from 1.54 (for a sample of 2.32 million articles and reviews published in 2009 or 2014) to 1.80 (for a sample of 69 261 highly-cited documents).

4.3.2 Coverage

After studying the size of GS, this section will focus on the characterisation of its content. To this end, its source, geographic, linguistic, discipline, and document type coverage will be discussed.

Source Coverage

Unfortunately, neither GS nor its secondary products (GSM and GSC) provide a master list of sources. As previously said, GS is not a journal database, but a service that indexes academic documents from many web domains. For this reason, efforts to determine its sources should try to identify these web domains first.

The first exhaustive study of the sources covered by GS based on web domains was carried out by *Aguillo* [4.32], who concluded that the most frequent geographic country code top-level domain (ccTLD) was ‘.cn’ (China), and that Harvard University was the higher education institution that contributed more content to GS. *Ortega* [4.3] went a step further by estimating the proportion of content provided by several types of content providers:

1. Publishers (41.6%)
2. Other Google products (22.5%)
3. Subject repositories (16.9%)
4. Institutional repositories (11.8%).

Martín-Martín et al. [4.37] analyzed the sources of the primary versions of a set of 64 000 highly-cited documents in GS, finding close to 6000 content providers, among which the US National Institutes of Health (nih.gov), ResearchGate (<http://researchgate.net>), and Harvard University (harvard.edu) were the main providers of highly-cited documents. This study also found that generic top-level domains (TLDs) like

Table 4.6 Compilation of self-elaborated materials on GS/WoS citation ratios at the document level

Source	Date of data collection	Description	N	GS citations	WoS citations	Ratio GS/WoS
Self-elaborated. Publication forthcoming	June–October 2016	Articles and reviews with a DOI covered by WoS, published in 2009 or 2014. WoS data extracted from web interface	2.32 million	42.6 million	27.6 million	1.54
Self-elaborated for this chapter	February 2017	Highly-cited documents in master sample. WoS data extracted only from GS/WoS integration	69 261	80.8 million	44.9 million	1.80

‘.edu’, ‘.org’, and ‘.com’ were more frequent than geographic TLDs. Lastly, *Jamali and Nabavi* [4.38], based on a series of topic queries, used GS to estimate the sources (<http://researchgate.net>, <http://nih.gov>) and top-level domains (‘.edu’ and ‘.org’) with a higher proportion of open-access documents.

Aiming to offer more updated results, we carried out a series of *site* queries in GS to find out the number hits returned for each of a list of 268 TLDs (251 geographic domains and 17 first-generation generic domains). The searches were carried out in March 2017. Publication year restrictions were not used, and cited references and patents were excluded.

Table 4.7 shows the main providers of documents according to their TLD. China is first among the geographic TLDs (12.12% of the total content), and commercial companies (.com) lead the list of generic domains (45.39% of the total content).

The sum of the results obtained for these 268 domains comes to 197 194 092 source documents, which is similar both to the one we obtained with the absurd query method in the previous section (184 001 450), and the one obtained by *Aguillo* [4.33], who used the same methodology (193 824 176). This reinforces the notion that the real number of source documents (excluding cited references) lies at around 200 million records. This figure is, however, a gross estimate, because there may be many duplicates, which will provoke an overestimation. At the same time, the fact that *site* command only counts primary versions would cause an infra-

estimation if the web domain of the primary version does not match the web domain queried with the *site* search command.

Otherwise, the results indicate that most of the content is hosted on websites with generic (not geographic) TLDs (69.7%), undoubtedly because of the weight of journal publishers, standalone journals, and American universities (.edu).

For the purpose of delving deeper into the issue of source typologies, we proceeded to calculate the size of five types of web domains. We wanted to illustrate the diversity and weight of the different types of sources from which GS feeds: digital libraries and bibliographic information systems (Table 4.8), publishers (Table 4.9), learned and professional societies (Table 4.10), US government agencies and international organizations (Table 4.11), and universities (Table 4.12). These tables present the number of results found for each element, both including and excluding cited references. The goal of these tables is to enable us to observe which sources generate a higher quantity of bibliographic records in GS. Lastly, it is worth remembering that these results only consider the primary versions of the documents (those GS has selected as primary versions), and, therefore, these tables should not be understood as a ranking of sources sorted by size, but rather, a list that shows the diversity of sources available in GS.

The results in Tables 4.8–4.12 illustrate the main sources from which GS feeds: big bibliographic information systems, including databases (Pubmed, Europe

Table 4.7 Top 20 domain sources of GS (2017)

Rank	TLD	%	HCE	Description	Type
1	.com	45.39	89 500 000	Commercial	Generic
2	.org	16.38	32 300 000	Noncommercial	Generic
3	.cn	12.12	23 900 000	China	Country
4	.edu	3.55	7 010 000	US accredited postsecondary institutions	Generic
5	.jp	3.40	6 700 000	Japan	Country
6	.net	2.06	4 070 000	Network services	Generic
7	.ru	1.72	3 400 000	Russian Federation	Country
8	.gov	1.69	3 340 000	US Government	Generic
9	.br	1.35	2 670 000	Brazil	Country
10	.fr	1.22	2 400 000	France	Country
11	.kr	0.94	1 850 000	Korea Republic of	Country
12	.ua	0.69	1 360 000	Ukraine	Country
13	.id	0.65	1 280 000	Indonesia	Country
14	.es	0.63	1 250 000	Spain	Country
15	.pl	0.56	1 110 000	Poland	Country
16	.de	0.51	1 010 000	Germany	Country
17	.au	0.44	864 000	Australia	Country
18	.uk	0.44	863 000	United Kingdom	Country
19	.it	0.40	797 000	Italy	Country
20	.ca	0.37	734 000	Canada	Country

Notes: Cited references excluded, HCE: Hit count estimate

Table 4.8 Digital libraries (DL) and bibliographic information systems (BIS) sources of GS

Rank	DL and BIS	URL	Source Type	Hits	
				Citations excluded	Citations included
1	China National Knowledge Infrastructure	http://cnki.com.cn	Database	17 600 000	19 600 000
2	Google books	http://books.google.com	Engine search	3 860 000	9 300 000
3	JSTOR	http://jstor.org	Digital library	2 920 000	4 680 000
4	Europe PubMed Central	http://europepmc.org	Subject repository	2 290 000	4 310 000
5	ResearchGate	http://researchgate.net	Social network	2 020 000	2 040 000
6	Proquest	http://proquest.com	Database	1 670 000	1 750 000
7	Astrophysics data system	http://adsabs.harvard.edu	Database	1 510 000	2 040 000
8	J-STAGE	http://jstage.jst.go.jp	E-journal aggregator	1 460 000	1 750 000
9	Pubmed	http://ncbi.nlm.nih.gov	Subject Repository	1 360 000	3 350 000
10	Cyberleninka	http://cyberleninka.ru	Digital library	1 150 000	1 200 000
11	CAB direct	http://cabdirect.org	Database	1 100 000	1 100 000
12	Refdoc	http://cat.inist.fr	Database	1 080 000	2 390 000
13	Academia.edu	http://academia.edu	Social network	1 020 000	1 030 000
14	CiteSeerX	http://citeseerx.ist.psu.edu	Search engine	1 010 000	997 000
15	ERIC	http://eric.ed.gov	Database	635 000	695 000
16	AGRIS	http://agris.fao.org	Database	537 000	3 620 000
17	Semantic scholar	http://semanticscholar.org	Search engine	526 000	527 000
18	EBSCO	http://ebSCOhost.com	Database	479 000	479 000
19	Dialnet	http://dialnet.unirioja.es	Bibliographic portal	458 000	2 280 000
20	ARXIV	http://arxiv.org	Subject repository	403 000	407 000

^aERIC: Education Resources Information Center**Table 4.9** Publisher sources of GS

Rank	Publishers	URL	Hits	
			Citations excluded	Citations included
1	^a Elsevier 2	http://sciencedirect.com	9 340 000	9 410 000
2	John Wiley and Sons	http://wiley.com	5 960 000	5 970 000
3	Springer	http://springer.com	5 590 000	5 770 000
4	Taylor and Francis	http://tandfonline.com	3 200 000	3 240 000
5	Sage	http://sagepub.com	1 370 000	1 560 000
6	Lippincott Williams and Wilkins	http://lww.com	1 240 000	1 240 000
7	Cambridge University Press	http://cambridge.org	1 100 000	1 270 000
8	^b Oxford University Press 2	http://oxfordjournals.org	951 000	1 240 000
9	Walter de Gruyter	http://degruyter.com	595 000	622 000
10	Nature Publishing Group	http://nature.com	428 000	458 000
11	Karger Publishers	http://karger.com	347 000	348 000
12	Chemical Abstracts Service	http://pubs.acs.org	325 000	325 000
13	BioMed Central	http://biomedcentral.com	279 000	279 000
14	Emerald	http://emeraldinsight.com	236 000	259 000
15	PLoS	http://journals.plos.org	203 000	203 000
16	World Scientific Publishing	http://worldscientific.com	168 000	170 000
17	Hindawi	http://hindawi.com	167 000	193 000
18	Elsevier 1	http://elsevier.com	108 000	192 000
19	Inderscience Publishers	http://inderscienceonline.com	82 500	82 500
20	Brill	http://booksandjournals.brillonline.com	68 100	122 000

^a This publisher owns another web domain (elsevier.com), in which we obtained 108 000 additional documents^b This publisher owns another web domain (oup.com), in which we obtained 4290 additional documents

Table 4.10 Learned and professional societies of GS

Rank	Learned and professional societies	URL	Hits	
			Citations excluded	Citations included
1	Institute of Electrical and Electronics Engineers (IEEE)	http://ieee.org	3 410 000	3 650 000
2	Institute of Physics (IOP)	http://iop.org	667 000	702 000
3	Royal Society of Chemistry (RSC)	http://rsc.org	470 000	476 000
4	Association for Computing Machinery (ACM)	http://acm.org	447 000	601 000
5	American Psychological Association (APA)	http://apa.org	406 000	448 000
6	American Chemical Society (ACS)	http://acs.org	326 000	327 000
7	American Society of Microbiology	http://asm.org	253 000	256 000
8	International Union of Crystallography	http://iucr.org	120 000	122 000
9	American Mathematical Society (AMS)	http://ams.org	93 500	112 000
10	American Meteorological Society (AMS)	http://ametsoc.org	59 100	66 300

Table 4.11 Government agencies and international organizations sources of GS

Rank	Government agencies and international organizations	URL	Hits	
			Citations excluded	Citations included
1	National Institute of Informatics (NII)	http://nii.ac.jp	2 960 000	13 300 000
2	Japan Science and Technology Agency (JST)	http://jst.go.jp	2 740 000	3 060 000
3	US National Institute of Health (NIH)	http://nih.gov	1 420 000	3 380 000
4	Institut de l'information scientifique et technique	http://inist.fr	1 100 000	2 420 000
5	US Department of Education	http://ed.gov	636 000	696 000
6	Office of Scientific and Technical Information	http://osti.gov	635 000	1 040 000
7	Food and Agricultural Organization (FAO)	http://fao.org	545 000	3 510 000
8	Defense Technical Information Center	http://dtic.mil	505 000	644 000
9	National Aeronautics and Space Administration (NASA)	http://nasa.gov	205 000	252 000
10	National Criminal Justice Reference Service	http://ncjrs.gov	120 000	124 000

Table 4.12 University sources of GS

Rank	Universities	URL	Hits	
			Citations excluded	Citations included
1	Harvard University	http://harvard.edu	1 410 000	2 170 000
2	Pennsylvania State University	http://psu.edu	1 030 000	1 080 000
3	Universidad de La Rioja	http://unirioja.es	442 000	2 280 000
4	University of Chicago	http://uchicago.edu	329 000	346 000
5	Johns Hopkins University	http://jhu.edu	324 000	340 000
6	Universidade de São Paulo USP	http://usp.br	155 000	197 000
7	Masarykova Univerzita v Brně	http://muni.cz	121 000	125 000
8	Universiteit van Amsterdam	http://uva.nl	105 000	108 000
9	Universidad Complutense de Madrid	http://ucm.es	105 000	356 000
10	Helsingin yliopisto	http://helsinki.fi	91 400	142 000

Pubmed Central, ADS), big commercial publishers (Elsevier, Springer and Wiley in particular), other academic search engines (Semantic Scholar, CiteSeer, ...), subject repositories (arXiv.org), social platforms (ResearchGate, Academia.edu), as well as Google's own book platform (Google Books). It additionally includes, research government agencies (like the Japanese National Institute of Informatics and the Japan Science and Technology Agency), professional associations (IEEE), and universities (Harvard University is still leads this

group). It should be noted that the results of some institutions (especially universities) are influenced by the existence of bibliographic products that are hosted within these universities' domains (Dialnet at the University of La Rioja, CiteSeerX at Pennsylvania State University, AGRIS in FAO, ERIC in the US Department of Education, etc.).

However, these results should be interpreted with caution, because the methodology used to collect the data has several limitations: All hit counts displayed

by GS (and Google, for that matter) are only approximations, not exact figures. What is more, the *site:* command is not exhaustive either; it only works with the primary versions of the documents in GS. GS implements a procedure to group together all the versions of a same document that may be available on the web [4.39]: subject and institutional repositories, the author's personal website, a social platform, and the official *version of record* available in a journal or on publisher's website. From all the versions found by the search engine, one of them (usually the publisher version, if there is one) is selected as the primary version. The rest of the versions can be found under the *All x versions* link available below each record.

This means that when a search containing the *site:* command is carried out, the system will only return the records in which the source of the primary version matches the searched source, even though there may be many more records from that source that have not been considered primary versions. If we focus on the case of ResearchGate, we can see that the 2 020 000 documents found (Table 4.8) are very far from the over 100 million documents the company claimed to cover in March 2017 [4.40]. This difference can be explained in part by the documents that are covered as secondary and not primary versions. For these reasons, the results in Tables 4.8–4.12 are most likely an underestimation of the real coverage, although it provides important clues as to the main sources it indexes.

Geographic Coverage

The geographic coverage of the documents covered by GS is also difficult to analyze, because the system is not designed to carry out searches based on authors' institutional affiliations. Therefore, like with the

source coverage, a possible, although biased, approach is to analyze the distribution of geographic domains. *Orduna-Malea* and *Delgado López-Cózar*, using this methodology [4.41], found that the domains for the United States, China, and Japan were the ones that yielded the highest hit counts estimates (HCE), which is consistent with the results obtained by *Aguillo* [4.32]. Table 4.13 shows the top ten geographic domains according to their HCE, and their evolution in the last 6 years [4.32, 41].

This strategy, however, is imprecise, because it does not consider generic TLDs like '.com' and '.org', which are precisely the ones that are most used. This explains why the United States are clearly under-represented, since a great proportion of '.com' domains belong to institutions from this country [4.42].

There are few studies on the geographic distribution of documents in GS, and those few that have been published focus on the geographical origin of the journals indexed in GSM, and the comparison of these journals with the ones covered by WoS and Scopus in specific disciplines like communication [4.43], nursing [4.44], and library and information science [4.45] (Fig. 4.4). GS and GSM seem to get closer to the actual distribution of scientific journals by country of publication than the other databases.

Similar results were found in a study of over 9000 arts, humanities, and social sciences journals indexed in GSM, where GSM not only covers journals from more countries, but the English-language bias is clearly less pronounced than in Scopus and WoS (Fig. 4.5).

Linguistic Coverage

It is also difficult to find out the language distribution of the documents covered by GS, because the

Table 4.13 Top ten geographic domains of GS

Country	TLD	Hit counts estimate			(%)
		2010	2013	2016	
China	.cn	7 520 000	30 700 000	23 900 000	12.12
USA	— ^a	N/A	16 019 000	10 943 500	5.55
Japan	.jp	1 720 000	10 400 000	6 700 000	3.40
Russia	.ru	995 000	N/A	3 400 000	1.72
Brazil	.br	1 440 000	2 320 000	2 670 000	1.35
France	.fr	2 820 000	4 210 000	2 400 000	1.22
South Korea	.kr	481 000	1 720 000	1 850 000	0.94
Ukraine	.ua	210 000	N/A	1 360 000	0.69
Indonesia	.id	N/A	N/A	1 280 000	0.65
Spain	.es	907 000	2 990 000	1 250 000	0.63
Poland	.pl	220 000	N/A	1 110 000	0.56

2010 data: from [4.32]

2013 data: from [4.41]

2016 data: self-elaborated for this chapter. In all cases, citations are excluded; N/A: not available

^a US data is obtained by merging the results from '.us', '.mil', '.edu', and '.gov' web domains

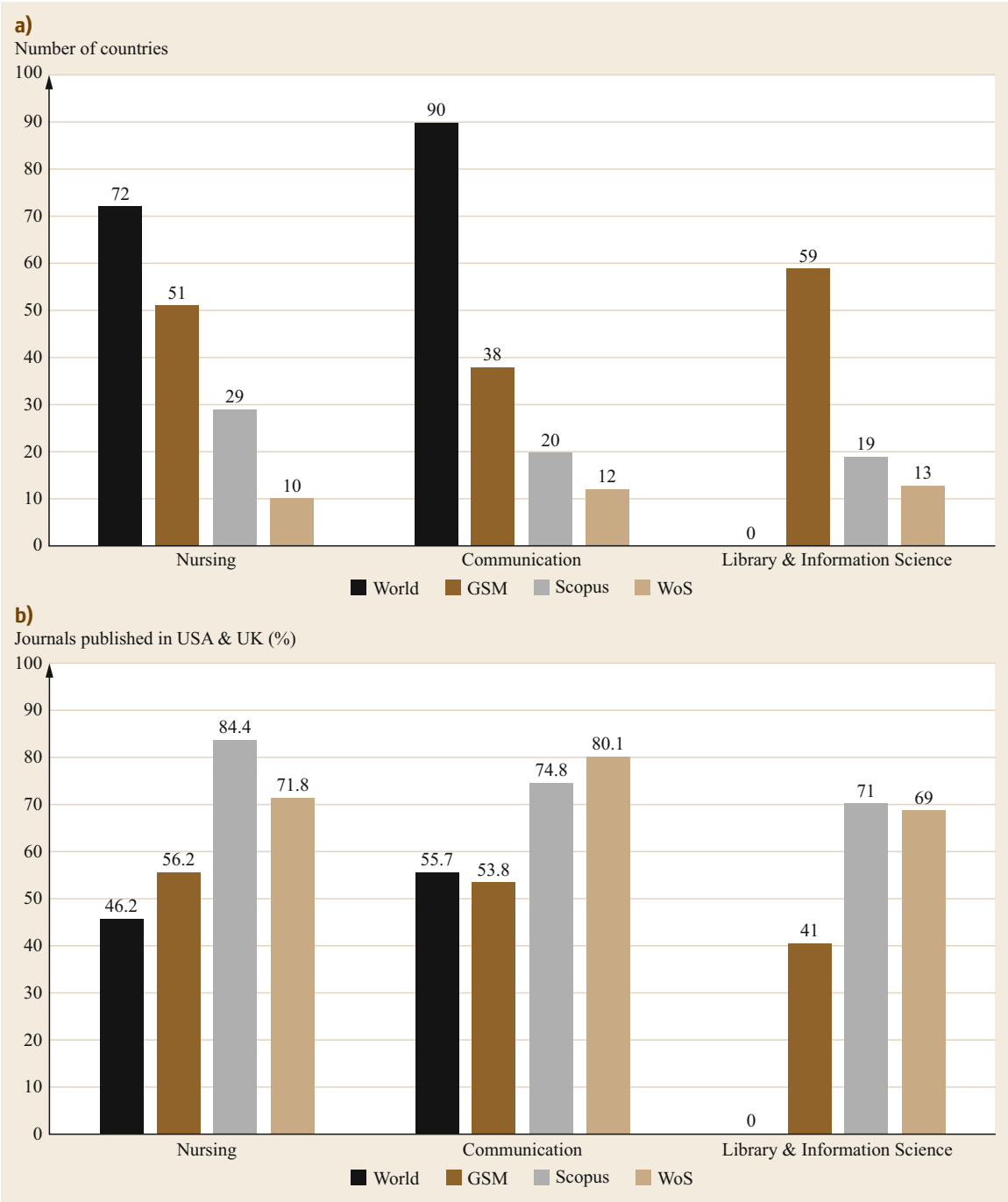


Fig. 4.4 Nursing, communication, and library and information science journals published according to GSM, Scopus, and WoS data (a). Number of countries where the journals are published (b). Percentage of journals that are published in the USA or UK. Note: In the case of library and information science journals, GS is used instead of GSM; and no data is available for the world category

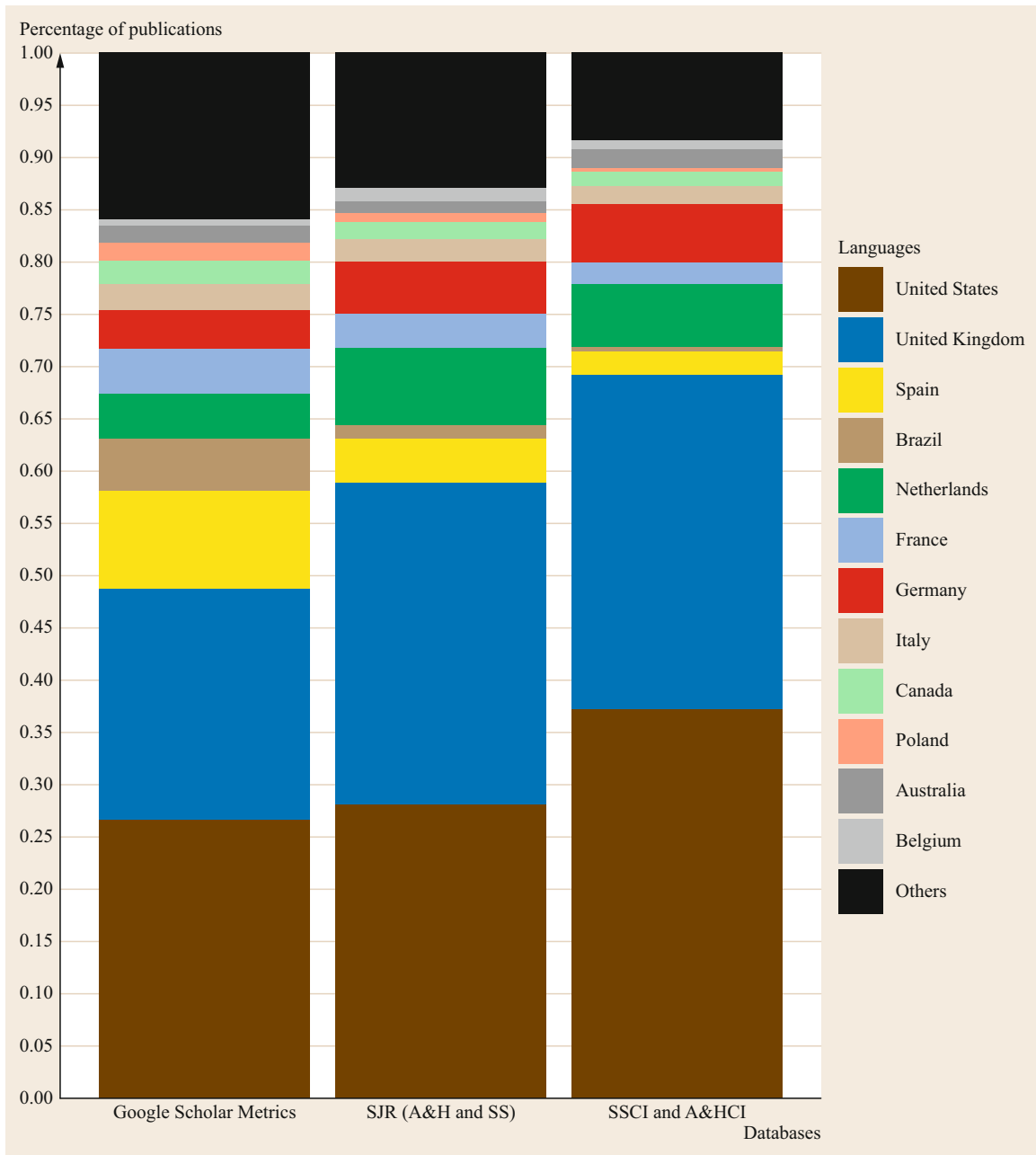


Fig. 4.5 Distribution of countries of publication in arts, humanities, and social sciences journals covered by GSM, SJR (Scopus), and JCR (WoS)

options the interface offers to filter by language are very limited. Users can limit search results to one or several of the following languages: Chinese (simplified and traditional), Dutch, English, French, German, Italian, Japanese, Korean, Polish, Portuguese, Spanish, and

Turkish. However, users have to navigate to the settings page to find these options; they are not available from the main search interface.

In a study carried out by *Orduna-Malea* and *Delgado López-Cózar* [4.41], which analyzed the quantity

of records by language in WoS, Scopus, and GS, a very high percentage of documents written in English was found in the first two databases (90%), while GS offered a higher linguistic diversity, because it covered other languages (especially Italian, Spanish, French, and Japanese). A year later, *Orduna-Malea et al.* [4.34], based on an analysis of the empirical studies on this topic, estimated that documents in the English language represented approximately 65% of all documents in GS.

Studies on the publication languages of journals available in GSM, as compared to those available in WoS or Scopus, reach the same conclusions. Both in communication journals [4.43], nursing journals [4.44], as well as in library and information science journals [4.45], GSM is closer to the actual distribution of languages used in scientific journals around the world, thus overcoming the bias towards English-language sources that prevails in Scopus and WoS (Table 4.14). While in the latter two, the proportion of English-language sources ranges from 80 to 93% of all sources, in GSM this figure is much lower: between 61 and 65% (Table 4.15).

An analysis of the arts, humanities, and social sciences journals available in the 2010–2014 edition of GSM yields similar results: GSM not only covers journals written in more languages, but its English-language bias is also lower than in the other two databases (Fig. 4.6), in spite of the fact that the study focused only on journals with titles written in Latin characters. GSM, because of its inclusion criteria, does not cover many journals for which articles are available in GS.

Aiming to obtain additional empirical data about the linguistic distribution of the content available in GS, we carried out a series of searches in GS. For each of the 12 languages that GS allows users to choose from to limit the search results (simplified Chinese, traditional Chinese, Dutch, English, French, German, Italian, Japanese, Korean, Polish, Portuguese, Spanish, and Turkish), 67 keyword-free, publication year queries

were carried out, one for each publication year for the period 1950–2016 (871 queries in total).

Table 4.16 shows the distribution of results by language (excluding cited references and patents). As can be seen, English-language results make up half of the total amount of results (49.8%), followed by the sum of simplified and traditional Chinese results (33.7%). These results are consistent with the figures on geographic coverage through the analysis of web domains presented previously and confirm the preeminence of the United States and China in GS’s coverage.

Even considering the disproportionately huge amount of English and Chinese results (which are clearly influenced by the sources from which GS extracts data), the distribution of the other languages is unquestionably more varied than the distribution presented by other databases. Figure 4.7 shows the relative distribution of languages in GS, Scopus, and WoS.

While the percentage of documents published in English in WoS and Scopus is of 90 and 80%, respectively, in GS the percentage is closer to 50%, and therefore the rest of languages are noticeably better represented.

Again, we would like to warn that the results obtained must be interpreted in the context of a search engine, and not a bibliographic database. GS automatically identifies the language of a document from certain parameters. However, a document might contain text in several languages. Therefore, the same document might be classified in various languages. Additionally, in some cases, the fact that a document is hosted in a geographic domain can help GS identify its language (for example, ‘.cn’ is associated with the Chinese language), even if sometimes the documents are not written in the expected language [4.46]. For those reasons, the number of results might be an overestimation.

Table 4.14 Number of different languages in nursing and communication journals indexed in GSM, Scopus, and WoS

Discipline	World	GSM	Scopus	WoS
Nursing	33	20	13	6
Communication	23	14	7	6

Table 4.15 Percentage of journals published in English in nursing and communication, and indexed in GSM, Scopus, and WoS

Discipline	World	GSM	Scopus	WoS
Nursing	57.2	61.9	81.2	92.7
Communication	70	65.3	91.6	87.8

Table 4.16 Distribution of languages in GS results (cited references and patents excluded)

Language	Documents	(%)
English	90 932 140	49.76
Chinese	61 545 203	33.70
Japanese	6 327 073	3.46
German	4 326 244	2.37
Spanish	4 144 354	2.27
French	3 657 705	2.00
Portuguese	2 403 898	1.32
Korean	2 131 744	1.17
Italian	999 134	0.55
Polish	766 266	0.42
Dutch	475 703	0.26
Turkish	472 830	0.26
Other	4 534 156	2.48
Total	182 716 450	100

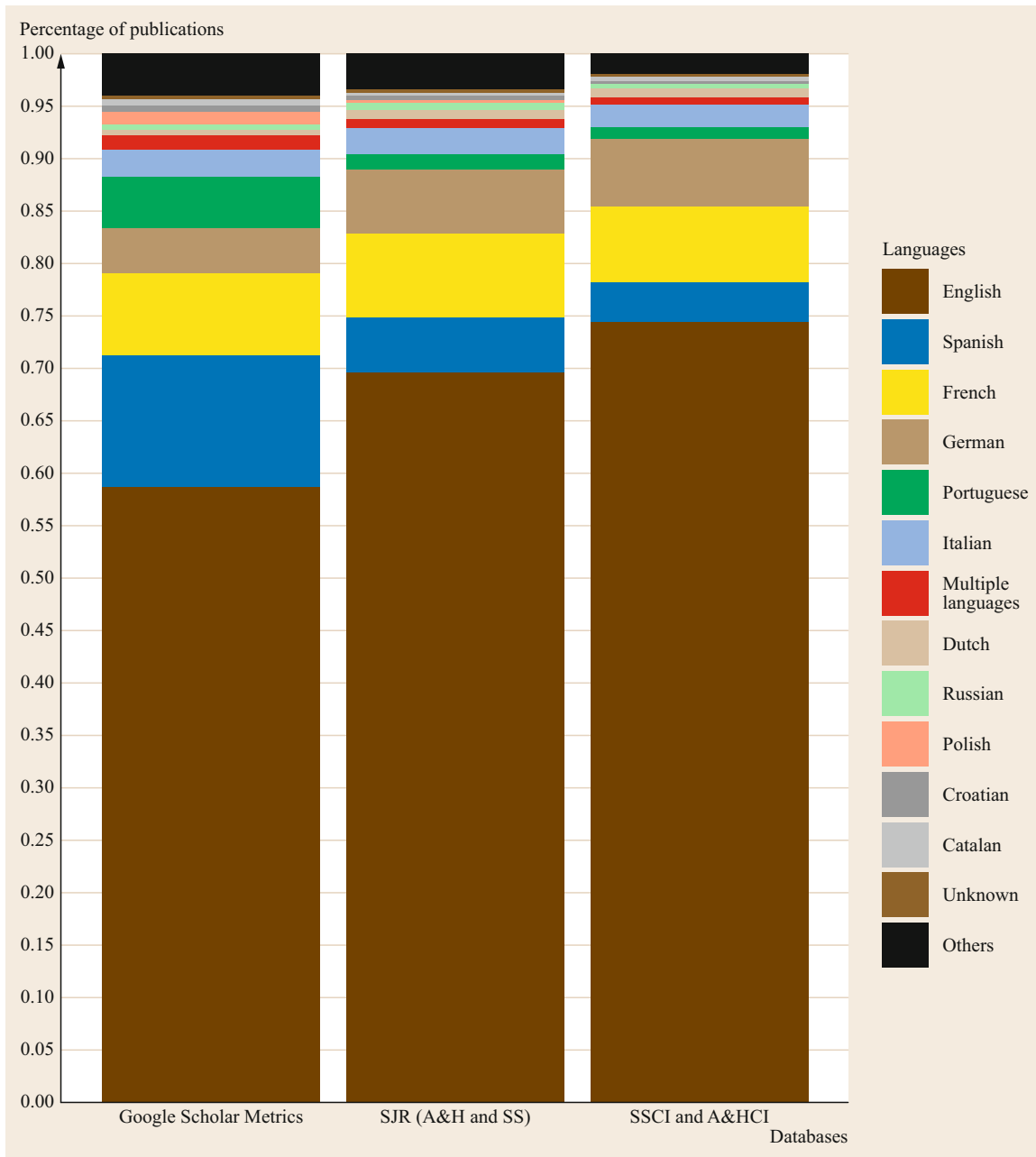


Fig. 4.6 Distribution of arts, humanities, and social sciences journals indexed in GSM (2010–2014) as compared to SCImago journal Rank (arts and humanities (A&H) and social sciences (SS) only), and journal citation reports (SSCI and A&HCI (Arts and Humanities Citation Index))

Discipline Coverage

Discipline coverage is another crucial aspect of the analysis of a bibliographic database. However, studies published to date mostly deal with journal coverage in GS as regards specific disciplines and countries.

From the data available about Spanish journals in the areas of social sciences, covered by GS and GSM

in the year 2011 [4.47], and the data available in IN-RECS [4.48], a ranking of Spanish journals in disciplines related to the social sciences, a very informative table about the coverage of GS and GSM was developed (Table 4.17). Of the 1090 Spanish journals studied in IN-RECS, 95.2% (1038) were covered by GS. What is more, for some disciplines, GS covered even more

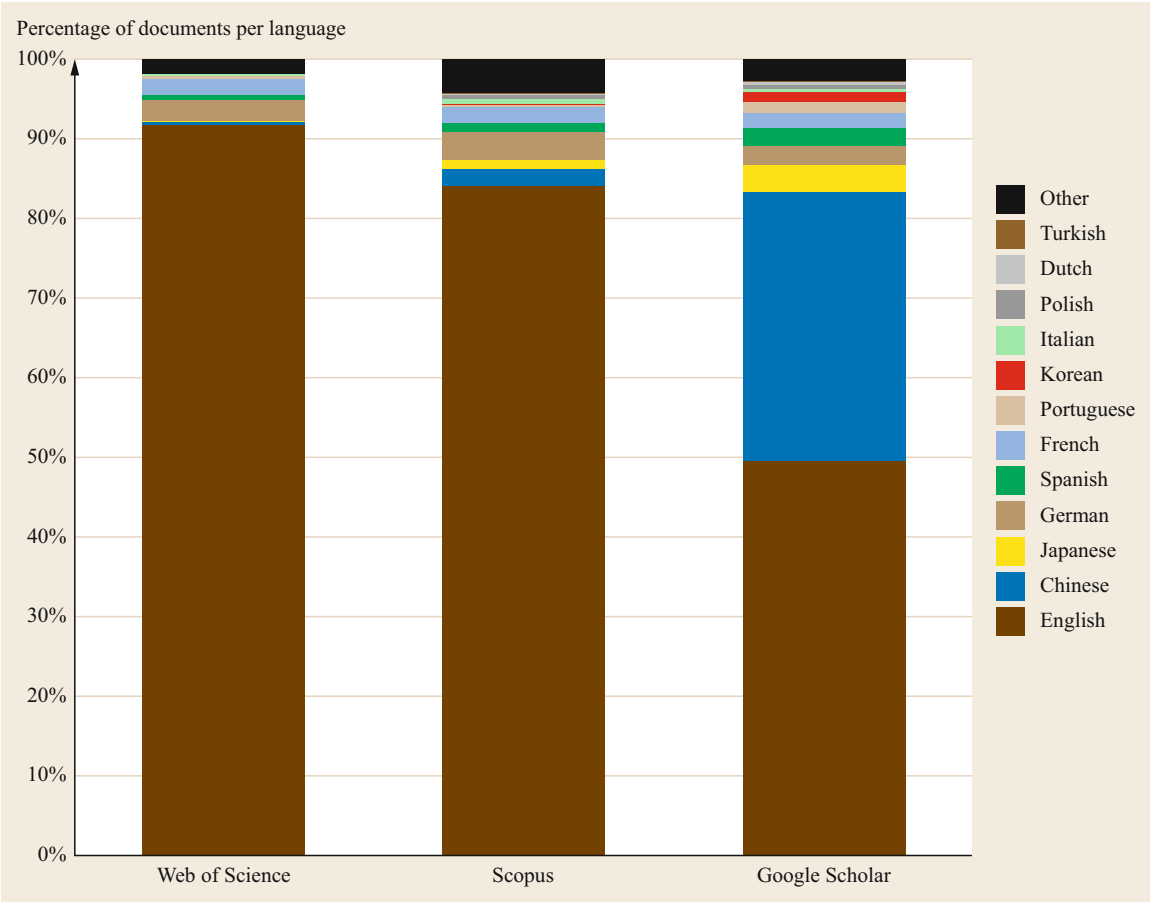


Fig. 4.7 Distribution of the languages of documents indexed in GS, Scopus, and WoS (1800–2016)

Table 4.17 Coverage of Spanish journals by discipline in IN-RECS, GS, and GSM

Discipline	IN-RECS	GS		GSM	
	Year (2010)	Year (2011)	Coverage (%)	Year (2011)	Coverage (%)
Law	341	251	74	110	43.8
Education	166	157	95	69	43.9
Economy	136	137	101	55	40.1
Psychology	108	109	101	42	38.5
Sociology	82	87	106	25	28.7
Political science and administration	60	56	93	21	37.5
Geography	51	54	106	15	27.8
Anthropology	46	46	100	10	21.7
Sport	N/A	42	N/A	14	33.3
Urban studies	43	39	91	15	38.5
Library and information science	33	36	109	15	41.7
Communication	24	24	100	12	50.0
Total	1090	1038	95	403	37.0

journals than IN-RECS. If we extrapolate those results, we can estimate that GS is very close to covering all active scientific journals. Of course, this hypothesis would require a more varied sample of journals to be tested.

On the other hand, GSM covers just over a third of all active Spanish social sciences journals (36.97%). This is most likely caused by the inclusion criteria enforced by the system (journals must have published

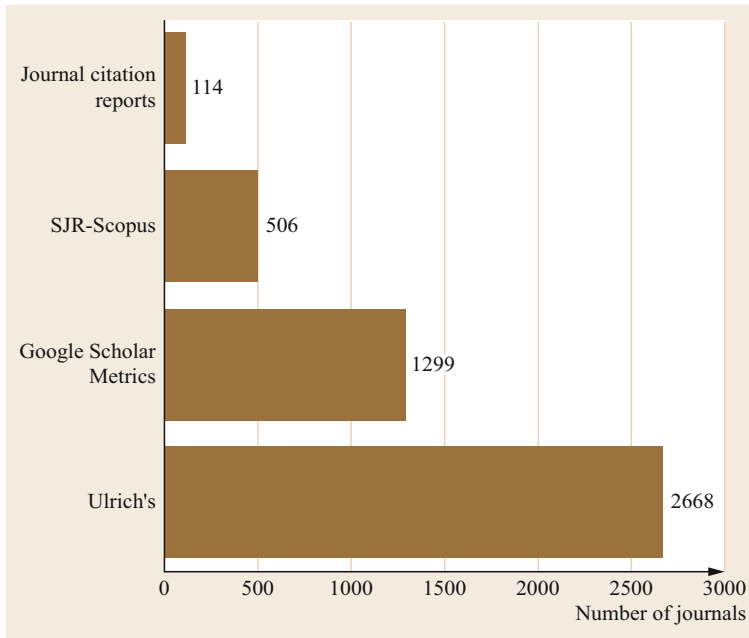


Fig. 4.8 Number of Spanish scientific journals covered by journal citation reports, Scopus, GSM, and Ulrich's directory

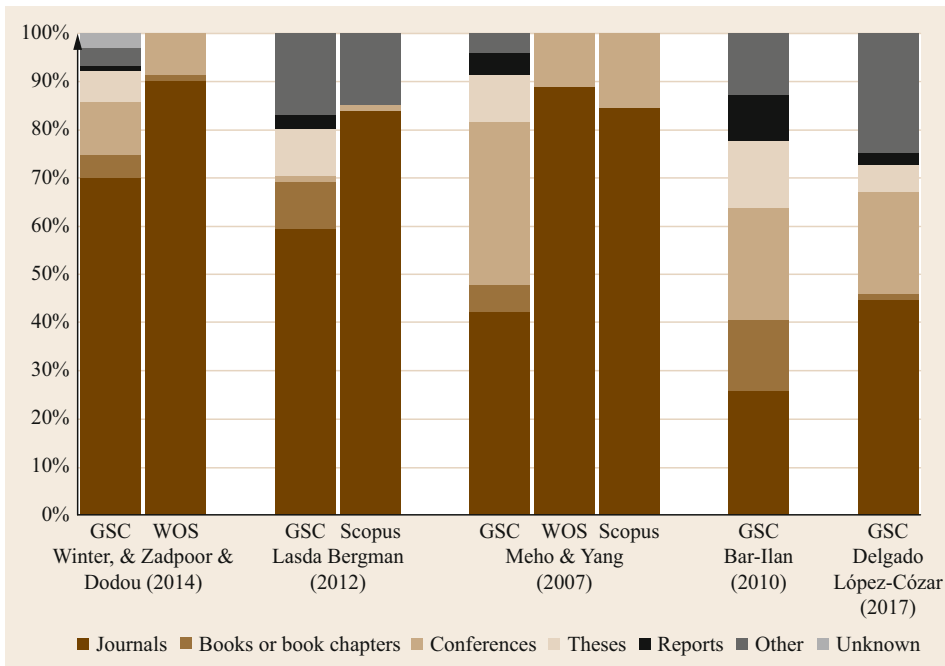


Fig. 4.9 Comparison of the document type distributions in GS, WoS, and Scopus as found in several studies

at least 100 articles in the last 5 years and have received at least one citation). Nevertheless, even with those limitations, GSM still covers many more journals than traditional databases (WoS and Scopus), and, therefore, is able to display a much broader spectrum of disciplines, better representing the scientific landscape.

In a study focused on the quantity of Spanish journals indexed in GSM, as compared with the total number of active Spanish journals, which according to the Ulrich's directory is around 2668 journals [4.49], similar results are obtained. Only 48.7% (1299) of Spanish journals were found in GSM (Fig. 4.8), but if this result is compared to the traditional journal rank-

ings, we find that, in spite of its inclusion criteria, GSM covers twice the amount of journals than SJR (Scopus data), and ten times more than the journal citations reports (WoS data). However, there are studies that yield different results. *Gu* and *Blackmore* [4.50] find that, of a sample of 41 787 refereed academic journals from all disciplines covered by the Ulrich’s directory, only 20.8% (10 354) were to be found in GSM, a lower amount than those that were found in SJR (32%, 15 911).

An additional method to quantify the number of publications covered by databases is to analyze the number of citations received by a sample of documents in a specific scientific discipline. With this approach, we can find studies focused on journals [4.51, 52] and researchers [4.53–56]. There is one other approach, based on the characterization of the documents returned by GS to sets of topic queries [4.38].

Important disciplinary differences can be observed in the list of studies that offer empirical evidences on the functioning of GS. While in the social sciences, the humanities, and engineering (especially in computer science) the ratio of citations in GS to citations in other sources is very high, the differences are much lower when STEM fields are analyzed. As was previously commented, the field of chemistry was initially covered poorly in GS, mostly due to the refusal of the American Chemical Society and other important publishers to be indexed by the search engine. These problems have already been solved [4.53, 57].

Document Type Coverage

Lastly, the last aspect of the coverage of GS has to do with document typologies. In order to learn about the wealth and diversity of document sources from which GS extracts data, we must summarize the results offered by the studies that have analyzed the distribution of citing documents according to their typology in several databases (Fig. 4.9). The main feature of GS is that it indexes more diverse document typologies. Indeed, it is the database where journal articles constitute a lower percentage of the total documents (from 28 to 70% depending on the samples). Conversely, in WoS and Scopus, journal articles make up 90% of the documents, which means there is a very limited coverage of conference communications and books.

Aiming to obtain more updated and representative results, a total of 871 queries (by language and year of publication) were carried out (including cited references and patents in the search), extracting the 1000 search results available for each query. There were 861 843 documents extracted [4.58].

Given that GS does not provide information about the typology of the documents that are displayed in

Table 4.18 Distribution of document typologies in GS (N = 861 839 documents)

Document type	TOT	EN	CN	JP	DE	ES	FR	PT	KO	IT	PL	NL	TR
Unknown	463 290	11 299	85 547	29 313	32 092	40 926	27 375	38 655	26 206	38 773	41 803	44 671	46 630
Article	260 211	39 264	38 550	34 416	16 087	10 384	21 357	15 998	33 606	7584	13 638	13 499	15 828
Book/Chapter	120 304	14 418	5637	1912	18 461	14 035	15 593	11 559	311	19 702	11 054	6233	1389
Thesis	10 919	33	3912	590	117	74	347	244	2891	35	35	702	1939
Conference	5741	718	162	447	159	375	626	505	240	587	468	439	1015
Other	649	20	112	305	2	3	62	2	93	2	0	1	47
Report	381	29	78	14	25	13	36	8	0	10	0	166	2
Unpublished	183	12	2	0	60	6	4	29	0	31	0	29	10
Patent	161	136	0	3	0	20	0	0	0	2	0	0	0
Total	861 839	65 929	134 000	67 000	67 003	65 836	65 400	67 000	63 347	66 726	66 998	65 740	66 860

EN: English; CN: Chinese; JP: Japanese; DE: German; ES: Spanish; FR: French; PT: Portuguese; KO: Korean; IT: Italian; PL: Polish; NL: Dutch; TR: Turkish

SERPs, document type identification becomes a rather complex task. By means of matching techniques to other data sources (WoS, CrossRef), and by applying a set of heuristics to extract more metadata from the websites that hosted the documents, the typology of 53.8% of the documents in the sample was ascertained (398 549). Table 4.18 shows the number of documents by typology and language.

If we focus only on the documents for which the typology could be identified, articles (65.3%) and books (30.2%) are the most frequent typologies, followed at a distance by doctoral theses and conference communications. However, we must bear in mind the limitations of the sample. The search strategy used (keyword-free searches which were only limited by year and language of publication), combined with GS's limitations (a maximum of 1000 results per query, sorted by GS's relevance ranking algorithm), result in a sample biased towards highly-cited documents, because the number of citations is the parameter that weighs the most in these types of queries [4.58].

In order to illustrate the bibliographic diversity in GS, Table 4.19 provides a list of the document types analyzed in the empirical studies that have addressed this issue.

Table 4.19 Document typologies found in GS

Bachelor's dissertations	Notes
Bibliographies	Presentation slides
Biographical items	Preprints
Blogs	Regulations
Book chapters	Reports
Book reviews	Research proposals
Books	Research reports
Civil service competitive examination reports	Reviews
Conference keynotes	Series
Conference paper proposals	Short survey
Conference papers	Student portfolios
Conference posters	Supplementary material
Conference presentations	Syllabi
Doctoral dissertation proposals	Term papers
Doctoral dissertations	Tweets
Doctoral qualifying examinations	Unpublished manuscripts
Editorials	Unpublished papers
Guidelines and clinical algorithms	Web documents
Interviews	Web pages
Journal articles	Working papers
Letters to the editor	Workshop papers
Master's theses	Yearbooks
Master's thesis proposals	

Clearly, GS's bibliographic wealth is due to the manner in which it indexes information: The search engine indexes any document that is hosted on the academic web, providing it meets certain technical and structural criteria. The consequence, at any rate, is that the presence of full text conference proceedings, book chapters, reports, patents, presentation slides (either from university courses, conferences, or other events), and especially monographs and doctoral theses make of GS a unique tool not only to find information, but also to find citation data that is not available anywhere else.

4.3.3 Growth Rate

Although Sect. 4.3.1 already presented some results as to this search engine's growth rate (even comparing it to Scopus and WoS), a sectional approach such as the one represented in Fig. 4.2 misses the main properties of GS, such as its dynamic nature. All content (both source documents and their citations, new or old) in GS is updated automatically.

As regards longitudinal analyses, *Harzing* [4.53, 59] studied the growth of citations to 20 Nobel Prize winners in 4 disciplines, detecting a growth of 4.6% from April 2012 to April 2013. A similar result (4.4%) was found by *Harzing* and *Alakangas* [4.54], where 146 senior researchers from the University of Melbourne were analyzed.

Retroactive growth (that is, the inclusion of documents published a long time ago) was addressed in part by *De Winter* et al. [4.52], who analyzed the relative difference between citation counts to a classic article up to mid-2005 measured in mid-2005 and citation counts up to mid-2005 measured in April 2013.

The speed with which GS indexes new source documents (and finds more citations to documents already in its document base) was addressed by *Moed* et al. [4.56]. The authors compute the indexing speed of GS for 12 journals in 6 different disciplines, and compare the results to those found in Scopus. Although there are differences among disciplines, and results are affected by the open-access policies of big publishers, the authors find that "the median difference in delay between GS and Scopus of indexing documents in Scopus-covered journals is about 2 months." The latest study on this issue to date was written by *Thelwall* and *Kousha* [4.60], which focuses on early citations to journal articles. They selected a sample of articles published in library and information science (LIS) journals between January 2016 and March 2017. The results in GS are compared to those in WoS, Scopus, and ResearchGate. The results in this study show that GS clearly outperforms all the other databases in terms of finding early citations,

Table 4.20 Speed of indexing for JASIST articles in GS

Article	Online publication	Scopus Index	WoS Index	GS				
				Index	Other version	Days since index	Online age	Index speed
1	20-Jan	Yes	No	No	Yes	–	–	–
2	24-Jan	Yes	No	No	Yes	–	–	–
3	27-Jan	Yes	No	Yes	No	56	58	2
4	21-Feb	No ^a	No	Yes	No	31	33	2
5	21-Feb	Yes	No	Yes	No	31	33	2
6	27-Feb	Yes	No	No	Yes	–	27	–
7	27-Feb	Yes	No	Yes	No	26	27	1
8	27-Feb	No ^a	No	Yes	No	26	27	1
9	27-Feb	Yes	No	Yes	No	26	27	1
10	27-Feb	Yes	No	Yes	Yes	N/A	27	–
11	07-Mar	Yes	No	Yes	No	17	19	2
12	07-Mar	Yes	No	No	Yes	–	19	–
13	13-Mar	Yes	No	No	Yes	–	13	–
14	20-Mar	No	No	Yes	No	N/A	6	–
15	20-Mar	No	No	No	Yes	–	6	–
16	20-Mar	No	No	Yes	No	3	6	3
17	20-Mar	No	No	Yes	No	3	6	3
18	20-Mar	No	No	Yes	Yes	N/A	6	–
19	20-Mar	No	No	Yes	No	3	6	3
20	20-Mar	No	No	No	Yes	–	6	–
21	20-Mar	No	No	No	Yes	–	6	–
22	25-Mar	No	No	No	Yes	N/A	1	–
23	25-Mar	No	No	No	No	–	1	–
24	25-Mar	No	No	No	No	–	1	–

^a Not indexed in Scopus because they are book reviews
N/A: Not available

although ResearchGate's data is quickly becoming an interesting source for citation data as well.

What follows is a small-scale analysis that aims to illustrate this phenomenon, which has consequences not only for document searches, but on the speed with which citations are detected in the system. Articles accepted by the *Journal of the Association for Information Science and Technology* (JASIST) and made available as advance online publications between January 1st and March 25th, 2017 were identified, noting the specific date when they were made available online. Secondly, those articles were searched in GS, and, in the cases when they were found, the exact date of indexing was saved (this information is available when documents are sorted by date). Knowing these two dates, we were able to compute the speed of indexing (number of days since the article was first available online, until GS picked it up). Results are displayed in Table 4.20.

On the date of data collection (March 26th) GS had indexed 13 out of 24 of the articles analyzed from the publisher's website, although in 4 cases the date of indexing in GS was not available, because these documents were previously available in GS as preprints. What is more, out of the 11 articles GS had still not picked up from the publisher's website, 9 were avail-

able from other sources (mainly subject or institutional repositories). Only the two most recent articles (available on the publisher's website only 1 day before the analysis was carried out) were not available in GS in any form. As regards the indexing speed, it ranges from 1 to 3 days. It is worth taking into account that there is a ± 2 day margin of error, because we know the date of indexing but not the exact hour. According to these data, it seems that it only takes around 2 days for documents published in JASIST to be indexed in GS, although a larger sample would need to be analyzed to confirm this for sure.

If we compare these results to the coverage of these documents in other databases, we can observe that Scopus had indexed the documents that had been made available up to March 13th. Although the exact date of indexing in Scopus is unknown, it was necessarily below 13, a very respectable speed considering it is a controlled database. However, these documents are classified as *in press* in the platform, and Scopus does not compute citations to documents until they are formally published in a journal issue, something that can take months. WoS does not index documents until they have been formally published, and therefore does not cover any of the documents in the sample.

4.4 Google Scholar's Data for Scientometric Analyses

Lastly, this chapter would not be complete without addressing the limitations of this search engine as a source of data for bibliometric analyses.

It is important to differentiate between the limitations from which this platform suffers by design for a specific purpose, and the various kinds of errors that the search engine makes when it processes data from the academic web. Errors are deviations from the expected or normal functioning of the tool (like, for example, the existence of duplicate citations, versions of the same document that have not been merged, incorrect or incomplete attribution of authorship, etc.). Limitations, on the other hand, refer to the characteristics that can compromise the suitability of the tool for a specific purpose, especially if it is not the original purpose for which the tool was first developed (like, for example, using GS as a source of data for bibliometric analyses, instead of as a search tool).

4.4.1 Errors in Google Scholar

The studies that have been published on the topic of errors found in GS are rather disorganized and superficial. There are few empirical evidences, and they often lack proper systematic study backed by representative samples. Most of the time, only anecdotal evidence is presented, without addressing the important issue of the degree of pervasiveness of the errors (how often the errors occur throughout the document base). The results

in these studies are difficult to summarize and compare, and they become obsolete very quickly, because GS is constantly being updated and introduces improvements to its algorithms regularly.

Following in the footsteps of the numerous and sharp studies carried out by *Jacsó* [4.12, 31, 54–69], below we present a taxonomy of the types of errors made by GS. We propose to divide errors into four broad groups (search-related errors, parsing-related errors, matching errors, and source-related errors), as described in Table 4.21. These errors can affect bibliographic records (authorship, source or year of publication, etc.) and citations themselves. A more in-depth discussion of the errors that can be found in GS was recently published by *Orduna-Malea* et al. [4.36].

4.4.2 Google Scholar Limitations

After describing the most common type of errors in the database, this section describes the main limitations for the use of GS as a source for bibliometric studies and research performance evaluation. To this end, we have prepared three descriptive sheets listing the limitations of GS search, GSC, and GSM, because although some limitations are present in all three products, some of them are particular to only some of them.

Each descriptive sheet contains several sections:

1. Coverage
2. Search and results interface

Table 4.21 Types of errors in GS

Type of error	Description
Search-related	Those related to the process of searching information
Parsing-related	Related to the process of identifying and extracting bibliographic information about documents from websites or full texts (including cited references)
Matching	Those related to the process of identifying different versions of a same document in order to remove duplicates
Source-related	Which affect the links that lead to the source in which the document has been found

Table 4.22 GS descriptive sheet

GS
Coverage
Lack of transparency in its coverage ^a :
<ul style="list-style-type: none"> ● There is no public master list of the sources GS indexes (publishers, repositories, catalogues, bibliographic databases and repertoires, aggregators, ...). ● There is no public master list of journals indexed in the platform.
Non-scientific and non-academic documents are also occasionally covered: course syllabi, library guides, tweets, ...
There is no accurate method to estimate the size of GS.
Data is not stable. GS is dynamic and reflects the state of the academic web at a certain moment in time, ... The irregularity and unpredictability of GS's indexing speed may bias some bibliometric analyses if it is not taken into consideration.
Full text files that exceed 5 MB can be found on GS, but their full text will not be indexed (cited references will not be analyzed).
Easy to manipulate: Anyone can obtain a fully or partially fabricated document indexed in GS by uploading it to a university domain or public academic repository.

Table 4.22 (continued)

GS
Search and results interface
The advanced search form is limited to four search dimensions: keywords (with assisted Boolean operators, and the possibility to search only in the title of the document, or anywhere in the article), authors, source of publication (journal, conference, ...), and year of publication.
The number of records displayed in each results page is 10 by default. It can be increased to 20 on the settings page. In the past, however, it was possible to increase this number up to 100.
Only the first 1000 results for any query can be displayed. Similarly, even if a document has received more than 1000 citations, only the first 1000 can be displayed when clicking the <i>Cited by</i> link.
Results can only be sorted by relevance or by date of publication:
<ul style="list-style-type: none"> ● Relevance: This is the default method. Although the specific parameters that are taken into consideration for this sorting method have not been publicly disclosed, it has been found that the number of citations received by documents, as well as the language of the document in relation to the user's preferred language, both weigh heavily in the relevance sorting algorithm. ● Publication date: limits the search to documents published in the current year.
There are only three result filtering options once a search has been made:
<ul style="list-style-type: none"> ● By document type, limited to three categories: case laws, patents, and articles. The latter category includes journal articles, books or book chapters, conference proceedings, technical reports, theses ... ● By type of record: users are given the option to remove cited references (documents GS has only been able to find in the reference lists of other documents) from the search results. By default, cited references are included in the search results. ● By year: it is possible to limit results to documents published in a given year, or a range of years.
It does not offer any features to analyze results or compute bibliometric indicators.
Quality of the data
In each search result, <i>authors</i> are displayed in the second line: below the title and next to the source of publication (usually a journal), the date of publication, and the name of the publisher or web domain where the document was found. The space allocated to the author data in this line is limited (usually between 30 and 40 characters), and, therefore, only the first three or four authors can be displayed, depending on the length of their names. In this line, authors are mentioned only by the initials of their first and second names and their surname. If these authors have created a public GSC profile and verified it with their institutional email, their name will contain a link to their public GSC profile.
For more complete author data, users can click on the Cite button, and there, export the record to BibTeX (or another reference manager format). The BibTeX record will display the full name of up to 10 authors of the document.
No data regarding <i>institutional affiliation</i> of the authors of the documents is available (institution, country). Therefore, it is not possible to carry out studies on geographic and institutional production and collaboration.
There is no information available about the <i>language</i> in which documents are written, even though internally they must have this information, because users can choose to limit results to documents written in one or more of the following languages: simplified Chinese, traditional Chinese, Dutch, English, French, German, Italian, Japanese, Korean, Polish, Portuguese, Spanish, and Turkish.
The <i>typology of each document</i> is not clear (book, journal article, conference communication, thesis, report, ...). Only books are marked as such, usually when they have been found on Google Books.
Not all documents have an <i>abstract</i>
<i>Author-supplied keywords</i> are not available. The same happens with the descriptors used by the databases where the records are found.
The <i>list of cited references</i> in each article is not available either (even though they definitely have that information, since they need it to compute citations), making it difficult to carry out studies that require cited reference analysis.
Errors in the parsing routine can lead to numerous problems. There still is no a conclusive study about the type and degree of occurrence of these errors, but among them we can find:
<ul style="list-style-type: none"> ● Poor bibliographic description of documents: incorrect or missing titles, authors, source of publication, date of publication, ... ● Duplicate records, when GS is not able to match two or more records that actually refer to the same document. This can also lead to split citation counts, since some of the citations will be attributed to one of the versions of the document, and some to the other versions.
GS does not rely in any kind of controlled vocabulary for author names, journals, publishers, institutions, ... that facilitates the identification of the different name variants for these entities.
These last two limitations make it more difficult to carry out large-scale studies using GS data, since the data would have to go through important cleaning and normalization processes prior to the analysis.

Table 4.22 (continued)

GS
Data reuse and exporting capabilities
Users can copy citations for single records, in a variety of formats, by clicking on the <i>Cite</i> button below every record.
Records can be exported to reference managers manually one by one, also by clicking on the <i>Cite</i> button. The available formats are BibTeX, EndNote, RefMan, and RefWorks. Alternatively, users can also save records to the <i>My library</i> feature by clicking on the <i>Save</i> button (for which it is necessary to be logged in to a Gmail account). <i>My library</i> allows users to export up to 20 records in one go, in BibTeX, EndNote, RefMan, or CSV format. The abstract is never included as part of the exported records.
GS does not offer, nor is it planned to offer (at least in the near future), any kind of public application programming interface (API) to enable users to access and export data from GS in bulk.
A strict CAPTCHA system is in place to discourage users from making too many queries to the platform too quickly. Users (or bots) that go over a certain (undisclosed) number of queries in a certain time are asked to solve a CAPTCHA of some sort before they are able to continue their searches. Sometimes, if the system detects too many searches made from the same IP, that IP can get blocked temporarily.
^a This limitation also affects GS citations and GSM

3. Quality of the data
4. Data reuse and exporting capabilities.

The information included has been taken from the authors' own observations and empirical tests. Due to

length of the sheets, the sheet that describes the limitations of GS is given in Table 4.22, and the tables that list the limitations of GS citations and GSM can be found in the electronic supplementary material [4.14, see Appendix VI].

4.5 The Expanded Academic World of Google Scholar

The results of the analysis of the size, coverage, growth rate, and speed of indexing of GS fully justify considering this platform as a big data bibliographic tool (Fig. 4.10).

The empirical evidence described throughout this chapter allow us to affirm that GS is an all-inclusive tool, capable of bringing together not only the scientific world *stricto sensu* (that which is represented by WoS and Scopus as well), but the entire academic and

professional world in a broad sense, thus providing a much broader picture of academic activity [4.37]. Its coverage is the most well balanced of the commonly used multidisciplinary databases in terms of countries of publication, languages (no English bias), and document typologies (not only scientific articles), something which is crucial when analyzing fields where it is common to use channels of communication other than journal articles published in English, such as disciplines

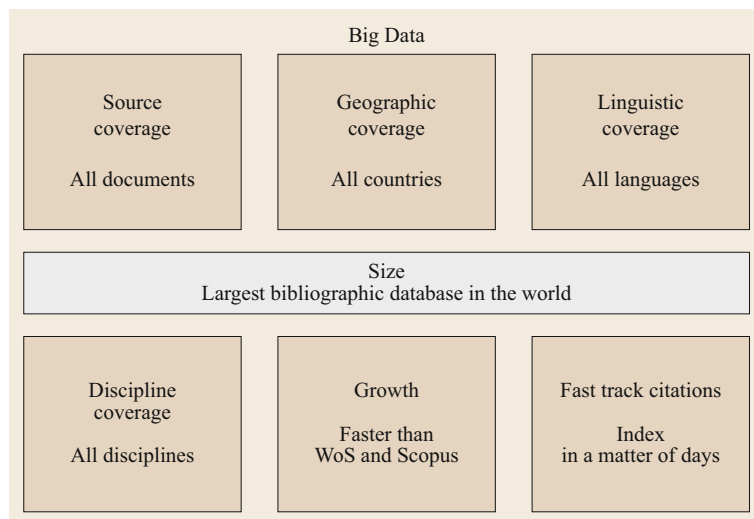


Fig. 4.10 GS, a big data bibliometric tool

in the arts, humanities, social sciences, and engineering. In this sense, Fig. 4.10 tries to convey the idea that GS covers all knowledge territories and all communities (scientific, educational, and professional), while WoS and Scopus only deal with scientific knowledge in the strict sense, and its communities.

The most important change over the previous paradigm, however, is GS's inclusion policy. WoS and Scopus have very exclusive and restrictive source inclusion policies. These sources are usually journals, which also place the prospective manuscripts researchers send to them under a rigorous evaluation processes (peer review). Contrary to this traditional model, GS works in a completely automated manner, without exercising any kind of selection process based on quality. Curated content from traditional journals and studies that have not gone through any kind of screening both coexist. GS automatically crawls, finds, and indexes any document that follows an academic structure and is hosted in an academic domain, even if it has not undergone any external quality control and is there only by decision of its authors. This breaks completely from the traditional controls to which all academic content had to be subjected prior to its public dissemination (peer review). For better or for worse, this is the distinguishing feature of GS: its ability to bring together reviewed and non-reviewed content, scientific and academic content (Fig. 4.11).

It should be pointed out that one of the main features shown in Fig. 4.11 is GS's nature as a receptacle. All the prestigious publications covered by WoS and Scopus are also covered by GS. When we observe the sources from which GS feeds, we can be sure, based on the empirical evidence on its size and coverage presented in the previous sections, that all the content covered by WoS and Scopus is also covered by GS. But of course, apart from the scientific elite, GS covers many other sources. We do not dispute that some of them may be of a lower quality, but others are of the same quality, if not higher, especially in certain disciplines (mainly doctoral theses, conference articles, working papers, and books).

Although it is true that GS covers sources that have not gone through a validation process like peer review (keynote conferences, syllabi, book reviews, technical reports, ...), these sources provide evidence of other kinds of impact beside the scientific impact, and could help put in a new light the work of researchers whose work is relevant to these communities and not the ones who publish in traditional databases.

Of course, the mixture of all these source documents (especially when considering that all of them are considered for computing citations) has been the object of important discussions in the bibliometric community [4.4]. To date, the main method to validate

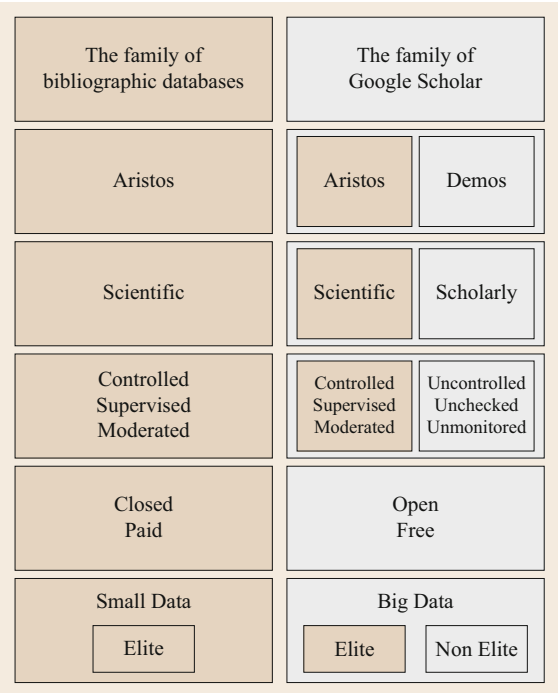


Fig. 4.11 GS versus traditional bibliographic databases

measurements made with data from GS has been to calculate their correlation with other well-established indicators. Many studies, which have been recently compiled by *Thelwall* and *Kousha* [4.70], have analyzed correlations between the number of citations according to GS and other databases (mainly WoS and Scopus), either at the level of journals or at the level of authors, as a way to evaluate its suitability as a source of data bibliometric studies.

The electronic supplementary material [4.14] offers a revised and updated list of these studies (see Appendixes II, III, and IV). Studies are grouped according to their unit of analysis (authors, citations, and average h-index). Although caution is advised for interpreting these results, because the nature of the samples are very different in terms of their size, time frames analyzed, and time when the studies were carried out, we can extract the following observations:

- The average correlation between GS and WoS (from 51 observations) is 0.76, and 0.81 when GS is compared to Scopus (28 observations).
- Out of all the studies, only in 2 are the correlations found to be below 0.50 (0.39 and 0.43 in Scopus, and 0.48 and 0.50 in WoS). All of them refer to correlations in the humanities and social sciences. On the other hand, there are numerous studies where the correlations found are above 0.9

(1 with GS/WoS comparisons and 7 with GS/Scopus comparisons).

- The highest correlations are found among STEM fields, and the lowest ones are usually found for fields in the humanities and social sciences.
- The differences between GS and the other databases seem to have increased in the more recent studies, which might indicate an even broader coverage in GS with respect to the other databases in recent years.
- No significative differences are found among studies with different units of analysis.

It also seems that multidisciplinary studies of international scope, and with very high sample sizes, achieve very high correlations, but that these become moderate by restricting the focus and emerging the intrinsic properties of each discipline. However, it should be borne in mind that correlations may have been calculated using different techniques (Pearson, Spearman, etc.), although values are reported independently, which may have a slight influence on the results.

To date, the largest sample studied for these purposes is the one used by *Martín-Martín* et al. [4.71], who used a sample of 64 000 highly-cited documents in

GS, of which 51% were also covered by WoS and had at least one citation. Most of the documents covered by both databases were journal articles, and the rest were monographs, theses, and conference articles. The R-squared (Spearman) found for the number of citations received by documents covered by both databases was $R^2 = 0.73$.

For this chapter, we decided to replicate the previous study, using the sample of 861 843 highly-cited documents in 13 languages (see Appendix V in the electronic supplementary material for further details about this dataset) [4.14, 58]. Out of these documents, 69 279 (8%) of them were covered by WoS and had received at least one citation. The Spearman correlation between the number of citations according to GS and according to WoS for these documents was $R_s = 0.9$. Figure 4.12 presents a scatter plot based on these data.

As Fig. 4.12 shows, the correlation between the number of citations these documents have received according to the two databases is evident. Additionally, an important number of observations seem to have received many more citations from GS than from WoS. This means that, even though the correlation is very high in general terms, GS is usually able to find many more citations than WoS for the same documents.

4.6 Final Remarks

GS is a prodigious mine of academic information that covers all fields of knowledge. Thanks to GS we now have access to previously unexplored territories of knowledge which, even if only roughly, allow us to form a broader mental picture of academic activity. This platform sheds light where previously there was only darkness. GS can definitely help to open the academic Pandora's box [4.4]. Opening this box will bring to light document typologies (especially monographs, theses, reports, conference communications, and book chapters) from the scientific core and periphery, which were previously invisible and inaccessible. The bibliographic features of these documents will probably change certain axioms and prejudices of academic evaluation. Previously undervalued researchers, clearly harmed (or outright forgotten) by the policies of traditional databases, will come to light, because there will be evidence of the impact of their work. Conversely, it will also confirm the poor performance of other researchers, until now protected by the lack of proper tools to evaluate them.

Perhaps all this will finally lead to a redesign of certain assessment and promotion systems, funding programs, and even research policies and structures. The

report *The metric tide* [4.72] already gives a glimpse of this changing trend from an institutional position, and not merely as an intellectual exercise advocated by a few and confined to research publications, with varying degrees of scientific impact, but without an actual practical impact.

That said, we cannot belittle the limitations of GS for bibliometric analyses. To begin with, the data exporting limitations (a maximum of 1000 results per query, and no easy way to export them). These obstacles are a hindrance when massive amounts of data are necessary for an analysis. The lack of an API in GS forces us to use third party applications (like Harzing's Publish or Perish) or download results manually. This results in very slow and costly data collection processes, which must be followed by a thorough cleaning of the raw data [4.63, 73, 74].

Additionally, GS does not provide vital information in its records, like the institutional affiliation of the authors, the language, and the document types. Not to mention the difficulties that normalizing the bibliographic information collected from so many varied sources entails. On the other hand, we do not believe that the errors from which some records suffer are

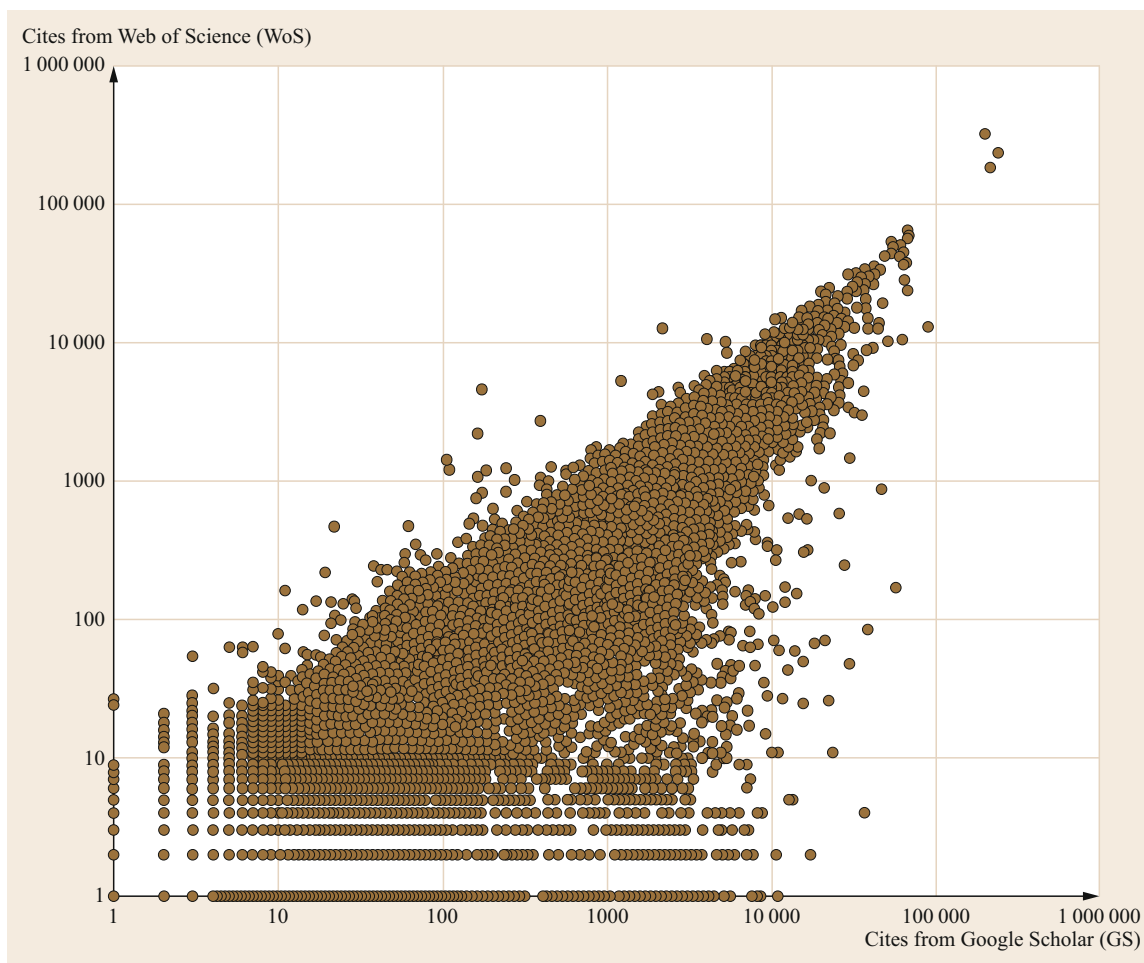


Fig. 4.12 Scatter plot of citations of highly-cited documents according to GS and WoS ($N = 69\,279$)

a major obstacle to the use of GS for bibliometric purposes. In a big data tool such as this, these errors are diluted and have no consequence on the big picture. These errors rarely affect individuals, journals, or other aggregates, as has been referred to by some studies already [4.75].

At any rate, the more dangerous limitations do not have to do with the methodological and technical problems previously discussed, but with the obscurity of the system, and, most of all, with the possibility of publication or citation manipulation, caused by the lack of quality control in the indexation of documents, which has been empirically proven by various studies [4.20, 76, 77]. Likewise, one of the main criticisms that is directed at GS is the lack of transparency, both regard-

ing its coverage (what sources it indexes) and updating mechanisms, and regarding its algorithm for ranking results after a query.

Lastly, we hope that this study will help readers understand the inner workings of GS and become aware of its enormous potential. We always tried to offer empirical evidence of its strengths (*contra data non argumenta*), without forgetting about the important dangers that its abuse could lead to. We hope to make the scientific community question assumed truths of an academic world of which only the tip of the iceberg has been visible until now. Let us explore its depths, let us observe and describe these new landscapes, and then let us decide if we would rather remain on the surface.

References

- 4.1 E. Garfield, I.H. Sher, R.J. Torpie: *The Use of Citation Data in Writing the History of Science* (Institute for Scientific Information, Philadelphia 1964)
- 4.2 E. Garfield: Science citation index: A new dimension in indexing, *Science* **144**(3619), 649–654 (1964)
- 4.3 J.L. Ortega: *Academic Search Engines: A Quantitative Outlook* (Elsevier, Oxford 2014)
- 4.4 E. Orduna-Malea, A. Martín-Martín, J.M. Ayllón, E. Delgado López-Cózar: *La revolución Google Scholar: abriendo la caja de Pandora académica* (UNE, Granada 2016)
- 4.5 Google Scholar: My citation counts have gone down. Help!, <https://scholar.google.com/intl/en/scholar/help.html#corrections> (2017)
- 4.6 S. Baggett, M. Williams: Budgets stretched, staff stressed, usage stalled... Something's gotta give! In: *Proc. Charleston Library Conf., Charleston*, ed. by B.R. Bernhardt, L.H. Hinds, K.P. Strauch (2012) pp. 324–327
- 4.7 B. Kramer, J. Bosman: *101 Innovations in Scholarly Communication—The Changing Research Workflow* (Universiteit Utrecht, Utrecht 2015)
- 4.8 S. Brown, S.A. Swan: Researchers' use of academic libraries and their services. A report commissioned by the Research Information Network and the Consortium of Research Libraries (2007)
- 4.9 J. Mussell, R. Croft: Discovery layers and the distance student: Online search habits of students, *J. Lib. Inf. Serv. Distance Learn.* **7**(1/2), 18–39 (2013)
- 4.10 D. Nicholas, C. Boukacem-Zeghmouri, B. Rodríguez-Bravo, J. Xu, A. Watkinson, A. Abrizah, E. Herman, M. Świgoń: Where and how early career researchers find scholarly information, *Learn. Publ.* **30**(1), 19–29 (2017)
- 4.11 R. Van Noorden: Online collaboration: Scientists and the social network, *Nature* **512**(7513), 126–169 (2014)
- 4.12 P. Jacsó: Google Scholar Author Citation Tracker: Is it too little, too late?, *Online Inf. Rev.* **36**(1), 126–141 (2012)
- 4.13 E. Delgado López-Cózar, A. Cabezas-Clavijo: Google Scholar Metrics: An unreliable tool for assessing scientific journals, *Prof. Inf.* **21**(4), 419–427 (2012)
- 4.14 E. Delgado López-Cózar, E. Orduna-Malea, A. Martín-Martín: Supplementary material to 'Google Scholar as a data source for research assessment', <https://doi.org/10.17605/OSF.IO/PQR53> (2017)
- 4.15 S.C. Bradford: Sources of information of specific subjects, *Engineering* **137**, 85–86 (1934)
- 4.16 Google Scholar: Google Scholar's inclusion guidelines for webmasters, <https://scholar.google.com/intl/en/scholar/inclusion.html#indexing> (2017)
- 4.17 Google Scholar: Google Scholar's About page, <https://scholar.google.com/intl/en/scholar/about.html> (2017)
- 4.18 Google Scholar: Google Scholar's search tips: Content coverage. What specific journals do you cover?, <https://scholar.google.com/intl/en/scholar/help.html#coverage> (2017)
- 4.19 E. Delgado López-Cózar, A. Cabezas-Clavijo: Ranking journals: Could Google Scholar Metrics be an alternative to journal citation reports and Scimago journal rank?, *Learn. Publ.* **26**(2), 101–113 (2013)
- 4.20 E. Delgado López-Cózar, N. Robinson-García: Repositories in Google Scholar Metrics or what is this document type doing in a place as such, *Cybermetrics* **16**(1) (2012)
- 4.21 A. Martín-Martín, J.M. Ayllón, E. Orduna-Malea, E. Delgado López-Cózar: Google Scholar Metrics 2014: A low cost bibliometric tool, *EC3 Working Papers* **17** (2014)
- 4.22 A. Martín-Martín, J.M. Ayllón, E. Orduna-Malea, E. Delgado López-Cózar: 2016 Google Scholar Metrics released: A matter of languages... and something else, *EC3 Working Papers* **22** (2016)
- 4.23 J. Bar-Ilan: Search engine results over time: A case study on search engine stability, *Cybermetrics* **2**(3) (1999)
- 4.24 R. Rousseau: Daily time series of common single word searches in Altavista and Northernlight, *Cybermetrics* **2**(3) (1999)
- 4.25 M. Thelwall: Quantitative comparisons of search engine results, *J. Am. Soc. Inform. Sci. Technol.* **59**(11), 1702–1710 (2008)
- 4.26 A. Uyar: Investigation of the accuracy of search engine hit counts, *J. Inform. Sci.* **35**(4), 469–480 (2009)
- 4.27 W. Koehler: Web page change and persistence—A four-year longitudinal study, *J. Am. Soc. Inform. Sci. Technol.* **53**(2), 162–171 (2002)
- 4.28 J.L. Ortega, I.F. Aguillo, J.A. Prieto: Longitudinal study of content and elements in the scientific web environment, *J. Inform. Sci.* **32**(4), 344–351 (2006)
- 4.29 N. Payne, M. Thelwall: Longitudinal trends in academic web links, *J. Inform. Sci.* **34**(1), 3–14 (2008)
- 4.30 A. Bosch, T. Bogers, T.M. Kunder: Estimating search engine index size variability: A 9-year longitudinal study, *Scientometrics* **107**(2), 839–856 (2016)
- 4.31 P. Jacsó: As we may search—comparison of major features of the Web of Science, Scopus, and Google Scholar, *Curr. Sci.* **89**(9), 1537–1547 (2005)
- 4.32 I.F. Aguillo: Is Google Scholar useful for bibliometrics? A webometric analysis, *Scientometrics* **91**(2), 343–351 (2012)
- 4.33 I.F. Aguillo: *Why Google Scholar? An Empirical Comparison of the Main Evaluation Tools* (Google Scholar Day, Madrid 2017)
- 4.34 E. Orduna-Malea, J.M. Ayllón, A. Martín-Martín, E. Delgado López-Cózar: Methods for estimating the size of Google Scholar, *Scientometrics* **104**(3), 931–949 (2015)
- 4.35 M. Khabsa, C.L. Giles: The number of scholarly documents on the public web, *PLoS ONE* **9**(5), e93949 (2014)
- 4.36 E. Orduna-Malea, A. Martín-Martín, E. Delgado López-Cózar: Google Scholar as a source for scholarly evaluation: A bibliographic review of database errors, *Rev. Esp. Doc. Cient.* **40**(4), e185 (2017)

- 4.37 A. Martín-Martín, E. Orduna-Malea, J.M. Ayllón, E. Delgado López-Cózar: A two-sided academic landscape: Snapshot of highly-cited documents in Google Scholar (1950–2013), *Rev. Esp. Doc. Cient.* **39**(4), e149 (2016)
- 4.38 H.R. Jamali, M. Nabavi: Open access and sources of full-text articles in Google Scholar in different subject fields, *Scientometrics* **105**(3), 1635–1651 (2015)
- 4.39 A. Verstak, A. Acharya: Identifying a primary version of a document, US Patent US8095876B1 (Washington DC 2013)
- 4.40 ResearchGate: About, <https://www.researchgate.net/about> (2017)
- 4.41 E. Orduna-Malea, E. Delgado López-Cózar: Google Scholar metrics evolution: An analysis according to languages, *Scientometrics* **98**(3), 2353–2367 (2014)
- 4.42 G.A. Barnett, C.J. Chung, H.W. Park: Uncovering transnational hyperlink patterns and web-mediated contents: A new approach based on cracking. com domain, *Soc. Sci. Comput. Rev.* **29**(3), 369–384 (2011)
- 4.43 E. Delgado López-Cózar, R. Repiso-Caballero: El impacto de las revistas de comunicación: Comparando Google Scholar metrics, Web of Science and Scopus, *Comunicar* **21**(41), 45–52 (2013)
- 4.44 L.M. Reina Leal, R. Repiso-Caballero, E. Delgado López-Cózar: H Index of scientific Nursing journals according to Google Scholar metrics (2007–2011), *EC3 Reports* 5 (2013)
- 4.45 E. Delgado López-Cózar, E. Orduna-Malea, D.M. Cartagena, E. Jiménez-Contreras, R. Ruiz-Pérez: Journal Scholar: Una alternativa internacional, gratuita y de libre acceso para medir el impacto de las revistas de Arte, Humanidades y Ciencias Sociales, *EC3 Working Papers* 5 (2012)
- 4.46 A. Martín-Martín, E. Orduna-Malea, A.W. Harzing, E. Delgado López-Cózar: Can we use Google Scholar to identify highly-cited documents?, *J. Inform.* **11**(1), 152–163 (2017)
- 4.47 E. Delgado López-Cózar, D.M. Cartagena, E. Jiménez-Contreras, R. Ruiz-Pérez: Índice H de las revistas españolas de Ciencias Sociales y Jurídicas según Google Scholar (2002–2011), *EC3 Reports* 4 (2013)
- 4.48 Grupo de Evaluación de la Ciencia y de la Comunicación Científica (EC3): IN-RECS database, <http://ec3.ugr.es/in-recs> (2017)
- 4.49 J.M. Ayllón, A. Martín-Martín, E. Orduna-Malea, E. Delgado López-Cózar: Índice H de las revistas científicas españolas según Google Scholar metrics (2011–2015), *EC3 Reports* 17 (2016)
- 4.50 X. Gu, K. Blackmore: Characterisation of academic journals in the digital age, *Scientometrics* **110**(3), 1333–1350 (2017)
- 4.51 K. Kousha, M. Thelwall: Google Scholar citations and Google Web/URL citations: A multi-discipline exploratory analysis, *J. Am. Soc. Inform. Sci. Technol.* **58**(7), 1055–1065 (2007)
- 4.52 J.D.F. De Winter, A.A. Zadpoor, D. Dodou: The expansion of Google Scholar versus Web of Science: A longitudinal study, *Scientometrics* **98**(2), 1547–1565 (2014)
- 4.53 A.W. Harzing: A preliminary test of Google Scholar as a source for citation data: A longitudinal study of Nobel Prize winners, *Scientometrics* **94**(3), 1057–1075 (2013)
- 4.54 A.W. Harzing, S. Alakangas: Google Scholar, Scopus and the Web of Science: A longitudinal and cross-disciplinary comparison, *Scientometrics* **106**(2), 787–804 (2016)
- 4.55 A.W. Harzing, S. Alakangas: Microsoft Academic: Is the phoenix getting wings?, *Scientometrics* **110**(1), 371–383 (2016)
- 4.56 H. Moed, J. Bar-Ilan, G. Halevi: A new methodology for comparing Google Scholar and Scopus, *J. Inform.* **10**(2), 533–551 (2016)
- 4.57 L. Bornmann, W. Marx, H. Schier, E. Rahm, A. Thor, H.D. Daniel: Convergent validity of bibliometric Google Scholar data in the field of chemistry citation counts for papers that were accepted by *Angewandte Chemie International Edition* or rejected but published elsewhere, using Google Scholar, Science Citation Index, Scopus, and Chemical Abstracts, *J. Inform.* **3**(1), 27–35 (2009)
- 4.58 A. Martín-Martín, E. Orduna-Malea, E. Delgado López-Cózar: Highly cited documents in Google Scholar by language and year of publication (January 2017), <https://doi.org/10.6084/m9.figshare.5024465.v1> (2017)
- 4.59 A.W. Harzing: A longitudinal study of Google Scholar coverage between 2012 and 2013, *Scientometrics* **98**(1), 565–575 (2014)
- 4.60 M. Thelwall, K. Kousha: ResearchGate versus Google Scholar: Which finds more early citations?, *Scientometrics* **112**(2), 1125–1131 (2017)
- 4.61 P. Jacsó: Google Scholar: The pros and the cons, *Online Inf. Rev.* **29**(2), 208–214 (2005)
- 4.62 P. Jacsó: Deflated, inflated and phantom citation counts, *Online Inf. Rev.* **30**(3), 297–309 (2006)
- 4.63 P. Jacsó: Google Scholar revisited, *Online Inf. Rev.* **32**(1), 102–114 (2008)
- 4.64 P. Jacsó: Testing the calculation of a realistic h-index in Google Scholar, Scopus and Web of Science for F.W. Lancaster, *Libr. Trends* **56**(4), 784–815 (2008)
- 4.65 P. Jacsó: The pros and cons of computing the h-index using Google Scholar, *Online Inf. Rev.* **32**(3), 437–452 (2008)
- 4.66 P. Jacsó: Calculating the h-index and other bibliometric and scientometric indicators from Google Scholar with the publish or perish software, *Online Inf. Rev.* **33**(6), 1189–1200 (2009)
- 4.67 P. Jacsó: Google Scholar's Ghost Authors, *Libr. J.* **134**(18), 26–27 (2009)
- 4.68 P. Jacsó: Metadata mega mess in Google Scholar, *Online Inf. Rev.* **34**(1), 175–191 (2010)
- 4.69 P. Jacsó: Google Scholar duped and deduped—The aura of 'robometrics', *Online Inf. Rev.* **35**(1), 154–160 (2011)
- 4.70 M. Thelwall, K. Kousha: Web indicators for research evaluation. Part 1: Citations and links to academic articles from the Web, *El profesional de la información* **24**(5), 587–606 (2015)
- 4.71 A. Martín-Martín, E. Orduna-Malea, J.M. Ayllón, E. Delgado López-Cózar: Does Google Scholar contain all highly cited documents (1950–2013)?, *EC3 Working Papers* 19 (2014)

- 4.72 J. Wilsdon, L. Allen, E. Belfiore, P. Campbell, S. Curry, S. Hill: *The Metric Tide: Independent Review of the Role of Metrics in Research Assessment and Management* (HEFCE, UK 2015)
- 4.73 L.I. Meho, K. Yang: Impact of data sources on citation counts and rankings of LIS faculty: Web of Science versus Scopus and Google Scholar, *J. Am. Soc. Inform. Sci. Technol.* **58**(13), 2105–2125 (2007)
- 4.74 J. Bar-Ilan: Citations to the “Introduction to informetrics” indexed by WOS, Scopus and Google Scholar, *Scientometrics* **82**(3), 495–506 (2010)
- 4.75 G. Doğan, İ. Şencan, Y. Tonta: Does dirty data affect google scholar citations?, *Proc. Assoc. Inform. Sci. Technol.* **53**(1), 1–4 (2016)
- 4.76 J. Beel, B. Gipp: Academic search engine spam and Google Scholar’s resilience against it, *J. Electron. Publ.* (2010), <https://doi.org/10.3998/3336451.0013.305>
- 4.77 C. Labbé, D. Labbé: Duplicate and fake publications in the scientific literature: How many SCiGen papers in computer science?, *Scientometrics* **94**(1), 379–396 (2013)

Emilio Delgado López-Cózar

Información y Documentación
Universidad de Granada
Granada, Spain
edelgado@ugr.es



Emilio Delgado López-Cózar is a Professor at the Facultad de Comunicación y Documentación of the Universidad de Granada. He holds a PhD in Library & Information Science. His research interests are focused on the analysis of the new sources of information and scientific evaluation (e.g., search engines like Google Scholar) and of the new metrics used to measure the visibility and the impact of scientific activity (Altmetrics).

Enrique Orduna-Malea

Audiovisual Communication,
Documentation, and History of Art
Universitat Politècnica de València
Valencia, Spain
enorma@upv.es



Enrique Orduna-Malea is Assistant Professor at the Polytechnic University of Valencia. He holds a PhD in Webometrics, a Master Degree in Management of Multichannel Contents, and MA in Library & Information Science. His research interests focus on the study of contents and users in the Web and the design of methods and models of analysis. He has co-authored 3 books, and 40+ articles included in JCR (Journal Citation Reports).

Alberto Martín Martín



Información y Documentación
Universidad de Granada
Granada, Spain
albertomartin@ugr.es

Alberto Martín Martín is a fourth-year PhD candidate at the University of Granada, thanks to a Doctoral Fellowship granted by the Ministerio de Educación, Cultura y Deporte in Spain. For his thesis project, he is studying the potential of the data available in Google Scholar for bibliometric analyses.