
ESTUDIOS / RESEARCH STUDIES

Google Scholar as a source for scholarly evaluation: a bibliographic review of database errors

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

* Universitat Politècnica de València

e-mail: enorma@upv.es | ORCID iD: <http://orcid.org/0000-0002-1989-8477>

** Facultad de Comunicación y Documentación. Universidad de Granada

e-mail: albertomartin@ugr.es | ORCID iD: <http://orcid.org/0000-0002-0360-186X>

e-mail: edelgado@ugr.es | ORCID iD: <http://orcid.org/0000-0002-8184-551X>

Received: 17-07-2017; Accepted: 08-09-2017.

Citation/Cómo citar este artículo: Orduna-Malea, E.; Martín-Martín, A.; Delgado López-Cózar, E. (2017). Google Scholar as a source for scholarly evaluation: a bibliographic review of database errors. *Revista Española de Documentación Científica*, 40(4): e185. doi: <http://dx.doi.org/10.3989/redc.2017.4.1500>

Abstract: Google Scholar (GS) is an academic search engine and discovery tool launched by Google (now Alphabet) in November 2004. The fact that GS provides the number of citations received by each article from all other indexed articles (regardless of their source) has led to its use in bibliometric analysis and academic assessment tasks, especially in social sciences and humanities. However, the existence of errors, sometimes of great magnitude, has provoked criticism from the academic community. The aim of this article is to carry out an exhaustive bibliographical review of all studies that provide either specific or incidental empirical evidence of the errors found in Google Scholar. The results indicate that the bibliographic corpus dedicated to errors in Google Scholar is still very limited ($n=49$), excessively fragmented, and diffuse; the findings have not been based on any systematic methodology or on units that are comparable to each other, so they cannot be quantified, or their impact analysed, with any precision. Certain limitations of the search engine itself (time required for data cleaning, limit on citations per search result and hits per query) may be the cause of this absence of empirical studies.

Keywords: Google Scholar; Academic search engines; Bibliographic databases; Errors; Quality.

Google Scholar como una fuente de evaluación científica: una revisión bibliográfica sobre errores de la base de datos

Resumen: Google Scholar es un motor de búsqueda académico y herramienta de descubrimiento lanzada por Google (ahora Alphabet) en noviembre de 2004. El hecho de que para cada registro bibliográfico se proporcione información acerca del número de citas recibidas por dicho registro desde el resto de registros indizados en el sistema (independientemente de su fuente) ha propiciado su utilización en análisis bibliométricos y en procesos de evaluación de la actividad académica, especialmente en Ciencias Sociales y Humanidades. No obstante, la existencia de errores, en ocasiones de gran magnitud, ha provocado su rechazo y crítica por una parte de la comunidad científica. Este trabajo pretende precisamente realizar una revisión bibliográfica exhaustiva de todos los estudios que de forma monográfica o colateral proporcionan alguna evidencia empírica sobre cuáles son los errores cometidos por Google Scholar (y productos derivados, como Google Scholar Metrics y Google Scholar Citations). Los resultados indican que el corpus bibliográfico dedicado a los errores en Google Scholar es todavía escaso ($n=49$), excesivamente fragmentado, disperso, con resultados obtenidos sin metodologías sistemáticas y en unidades no comparables entre sí, por lo que su cuantificación y su efecto real no pueden ser caracterizados con precisión. Ciertas limitaciones del propio buscador (tiempo requerido de limpieza de datos, límite de citas por registro y resultados por consulta) podrían ser la causa de esta ausencia de trabajos empíricos.

Palabras clave: Google Scholar; Motores de búsqueda académicos; Bases de datos bibliográficas; Errores; Calidad.

Copyright: © 2017 CSIC. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY) Spain 3.0.

1. INTRODUCTION

The launch of a new tool

Google Scholar (GS) is an academic search engine created by Google Inc. (now Alphabet) on 18 November 2004, and its main purpose is to provide “a simple way to broadly search for scholarly literature” and to help users to “find relevant work across the world of scholarly research”.¹

The way it functions is similar to the general Google search engine in that it is a system based on providing the best possible results to user queries entered into a stripped-down search box (Ortega, 2014). In the case of GS, it returns results for millions of academic documents (abstracts, articles, theses, books, book chapters, conference papers, technical reports or their drafts, pre-prints, post-prints, patents and court opinions) that the Google Scholar crawlers automatically locate in the academic web space: from academic publishers, universities, scientific and professional societies, to any website containing academic material (Orduna-Malea et al., 2016).

As with Google, the results retrieved for a particular query are ranked by an algorithm that takes into account a large number of variables (where it was published, who it was written by, how often and how recently it has been cited in other scholarly literature, etc.), although the exact components of this algorithm and the weight of each variable is unknown, for industrial property reasons. However, several empirical studies have demonstrated that the number of citations received by a document is one of the key ranking factors (Beel and Gipp, 2009; Martín-Martín et al., 2017). Another essential feature of Google Scholar is that the entire process is automated, without any human intervention, from the location of documents (crawling) to the bibliographic description (metadata parsing) and the extraction of the bibliographic references (reference parsing) that are used to compute the number of citations received by each retrieved document from all other documents.

Google Scholar was not the first tool of this type; other pioneering systems had already appeared on the scene (Citeseer, the first version of which dates from 1997, is considered the first academic search engine). However, the fact that it was developed under the umbrella of a company like Google, and used part of its technology, led to immediate acceptance by a significant proportion of the academic publishing world and by some professionals and researchers, a fact that was widely criticised by Jacsó (2006a), who openly mocked this new state of affairs (“As Google wandered into

the territory by launching Google Scholar (GS) at the end of 2004, the topic is expected to appear in the ultra-light morning television chat shows run by ultra-light TV personalities who are meant to light up our mornings”).

Given the characteristics of Google Scholar, it can and should be studied from two complementary but different angles (not only its characteristics but also its effects and consequences). Firstly, GS may be evaluated as a discovery tool (Breeding, 2015), that is, a search engine the purpose of which is to provide the best results to each query and a pleasant user experience based on usability, ease of use and, above all, speed (Bosman et al., 2006). Secondly, Google Scholar may be analysed as a tool that can be used to evaluate scholarly activity. This use, which came about due to it providing citation figures for each document indexed by the system, has led to the increasing use of Google Scholar by users (teachers, researchers, students) and professionals (companies, assessment bodies) as a bibliometric tool for various evaluation processes (authors, journals, universities), although it was not designed with this purpose in mind and lacked the required basic functions (Torres-Salinas et al., 2009). It is precisely this aspect (Google Scholar as a valid tool for carrying out bibliometric studies) that the objectives of this bibliographic review will be based on.

The launch of a new debate

The debate about the advantages and disadvantages of using Google Scholar began immediately after it first appeared (November 2004), giving rise to good and bad criticism in equal measure, as Giles (2005) pointed out in his column in *Nature*. The first analyses of Google Scholar came from technology blogs and websites, such as Sullivan’s (2004) more neutral and informative piece for *Search Engine Watch* (<https://searchenginewatch.com>), or Kennedy and Price’s (2004) more sensationalist piece for the now-defunct *Resource Shelf*, affirming that “as you’ve read here many times, Google is brilliant (that is, ingenious at marketing and trying new things), and this is yet another example of their savvy”. These messages propagated fast on the internet.

In spite of the general enthusiasm, critical voices soon made themselves heard, one of whom was Péter Jacsó (2004), who tested the search engine between 18 and 27 November 2004, publishing his findings informally on a blog.² In his study, Professor Jacsó, a specialist on database evaluation with extensive experience, conducted an analysis of the general coverage of various publishers

on Google Scholar using the "site" command, and identified a number of important limitations, leading him to conclude that "Google Scholar needs much refinement in collecting, filtering, processing and presenting this valuable data" (Jacsó, 2004). The issues identified by Jacsó included unfriendly search syntax, little or no information about the features of the search engine, and inconsistent results. He found specific errors, such as the way in which it displayed results in which there were changes in the word order of the title, or generated completely erroneous bibliographic descriptions (the book, *Computers and Intractability*, by Garey and Johnson, detected errors and inconsistencies in the title, subtitle, author names, publisher's names, locations and years). He also noted a wide range of additional errors like inflated hit counts, inflated citedness, full-text links pointing to erroneous documents and unmerged document versions.

At that precise moment, and in the wake of Jacso's criticism, a wave of criticism was directed against the general drawbacks of Google Scholar (Price, 2004; Goodman, 2004; Abram, 2005; Gardner and Eng, 2005; Notess, 2005; Ojala, 2005; Vine, 2005; Wleklinski, 2005; Adlington and Benda, 2006; White, 2006), alongside more neutral articles, such as the study published by Noruzi (2005), that, while acknowledging its many drawbacks, also pointed to its potential benefits and possible improvements. At the same time, other articles adopted a markedly neutral attitude towards GS. These included the columns by Butler (2004) in *Nature* and Leslie (2004) in *Science*, brief news features that did not discuss or even mention critical aspects, perhaps due in part to the fact that both the Nature Publishing Group and the American Association for the Advancement of Science (AAAS), the publishers of *Nature* and *Science*, respectively, had reached agreements to provide access to the full text of their publications to Google Scholar crawlers.

On the other hand, the paper published by Belew (2005) was a significant departure in the debate about the value of Google Scholar. This author analysed a corpus of 203 publications concluding that, surprisingly, there was a high correlation between the citations received by these documents according to Google Scholar and to ISI (the author did not indicate what exact database he used or the discipline to which the documents belonged, simply that six authors from the same interdisciplinary department had been chosen at random; in any case, the use of WoS may be surmised in the area of computer science). Similarly, Pauly and Stergiou (2005) conducted a citation analysis on a corpus of 114 articles from a

wide range of disciplines (mathematics, chemistry, physics, computing sciences, molecular biology, ecology, fisheries, oceanography, geosciences, economics, and psychology), and also observed a high correlation ($R^2 = 0.994$ for articles published from 2000 to 2004), which led them to affirm that "GS can substitute for ISI", and that "GS may gradually outperform ISI given its potentially broader base of citing articles". Finally, that same year, the seminal article by Bauer and Bakkalbasi (2005) appeared, an analysis published in D-Lib in which they compared "the citation counts provided by WoS, Scopus, and Google Scholar for articles from the *Journal of the American Society for Information Science and Technology (JASIST)* published in 1985 and in 2000". This study concluded that for articles published in 2000, Google Scholar provided statistically significant higher citation counts than either Web of Science or Scopus, and was significant because the authors brought to light the importance that citation analysis had acquired, not only for crawling academic publications or measuring their impact, but also for justifying tenure and funding decisions, underlining the future role that GS could play in this complex matter. Indeed, in the light of Bauer and Bakkalbasi's article, *The Scientist* devoted an article to the future of citation analysis and the role that the web in general and GS in particular could play in bibliometric analysis (Perkel, 2005).

Jacsó's response to these articles was not long in coming; he lambasted them in a column published in *Online Information Review* (Jacsó, 2006a). First, he declared his utter disagreement with Butler, claiming that he did not seem to have understood his illustrative examples of Google Scholar's errors, "even if my examples were as much tailor-made for *Nature* as bespoke suits by Savile Row tailors for the ultra rich". Second, he warned readers not to limit their reading to Belew's work. Third, with respect to Pauly and Stergiou, he openly criticised their claim that GS can replace ISI, particularly since their claim was arrived at by "handpicking" only a few articles, without filtering or even cleaning them up, and they contained numerous errors in the form of inflated and phantom citation counts. Two years later, Harzing and Van der Wal (2008) criticised Jacsó for his criticisms of Pauly and Stergiou in a seminal article published in the same journal in which these authors published their earlier article (*Ethics in Science and Environmental Politics*). They accused Jacsó of also "handpicking" examples of errors, with few and unrepresentative samples, and while they did acknowledge the errors pointed out by Jacsó, these errors were basically inconsistencies in the results for specific queries.

Lastly, Jacsó (2006a) acknowledged the validity of Bauer and Bakkalbasi's findings, although he recommended that readers take a critical look at the volume of citations not only in the 2000 sample (where GS was superior to WoS), but also in the 1985 sample (where WoS outperformed GS), data that seemed to have been overlooked by the academic community, which was more interested in highlighting only the positive aspects of GS and hiding or minimising its limitations, according to Jacsó.

From that moment, and in the same column in *Online Information Review* (called *Savvy Searching*), Péter Jacsó published a series of articles from 2006 to 2012 aimed at identifying, describing, categorising and denouncing the many errors and limitations of Google Scholar (listed in Appendix I, along with various data related to the errors identified and samples used in each study). Much of his research was also published on his personal website (www.jacso.info), as a way of archiving the evidence.

In spite of the strong and harsh criticism that he then fired off from his platform on *Online Information Review* (some of his most vehement remarks are listed in Table I), and which will be described in detail in the following sections, Jacsó was always rigorous, admitting that Google Scholar is an excellent tool for locating documents that might not be accessible through traditional databases, as well as for accessing full texts (i.e. as a discovery tool). However, "using it for bibliometric or scientometric purposes, such as for determining the h-index of a person or a journal, is another question" (Jacsó, 2008c). This fact led him to criticise colleagues that used Google Scholar for said purposes even if they did admit the limitations of the database. For example, Bar-Ilan (2008), in her study of highly-cited Israeli authors, admitted that "the sources and the validity of the citations in GS were not examined in this study". In the light of this observation, Jacsó (2008b) raised his dissenting voice, although he did qualify his position with an understanding that it is sometimes not only tedious but impossible to verify the origin and validity of the citations due to the system's significant limitations, laconically concluding, "I cannot blame her and others who accept the citation counts as reported by GS".

The debate came to a head in 2012 when a controversial article published by Jerome K. Vanclay (2012) in the journal *Scientometrics* strongly criticised the Impact Factor and advocated the use of alternative sources for the evaluation of journals, including Google Scholar. The controversy was heightened all the more by the tone employed by Vanclay. Therefore, Tibor Braun (the founder and editor-in-chief of *Scientometrics*

at the time) invited Jacsó to reply (Jacsó, 2012b). Jacsó's criticisms were extremely strong ("utterly demagogue rhetoric, featuring false accusations, misleading statements, claims and comparisons, delusional ideas, arrogance and ignorance in the Vanclay-set"), so much so that he even questioned the review and publication process ("part of a mock-up scenario to test how poorly researched, prejudiced, biased, duplicate papers using 'flawed methodology', ignorant arguments, erroneous calculations, loaded rhetoric, and misleading examples can get through the current quality filters of editorial preview and peer reviews").

The ideas of Vanclay (2012) were equally criticised by Butler (2011) and by Bensman (2012), who highlighted Vanclay's lack of knowledge about the workings and purposes of the WoS and JCR databases, and his excessive idealism in the "promise of a far better assessment of research/publication performance through the h-index based on GS".

Jacsó (2012b) once again reiterated that his criticism of Google Scholar was not directed towards its undoubted advantages for thematic searches, but towards its serious limitations, which make it inappropriate for bibliometric analysis ("extremely lenient citation matching algorithm"), an aspect with which Aguillo (2012) also concurred. In particular, Jacsó argued that the adulation shown by the bibliometric community towards Google Scholar is due in part to the fact that it retrieves a greater number of publications and citations and, consequently, a higher h-index than many researchers would deserve. This may have a perverse effect on the evaluation of the quantity and quality of publications in "decisions related to tenure, promotion and grant applications of individual researchers and research groups, as well as in journal subscriptions and cancellations" (Jacsó, 2012c).

The evolution of an – already old – debate

Between 2004 and 2008, criticism of Google Scholar was largely sustained by Jacsó's articles. However, other authors also expressed their reservations about this search engine, particularly because of the lack of improvements and updates, as Gregg Notess (2008) noted in the forum *Search Engine Showdown*.³ At around the same time, a report by the consulting firm comScore, published by the prestigious technology blog *Techcrunch*,⁴ reported a fall in the number of unique visitors to Google Scholar during the November 2006 to November 2007 period. This news was picked up on by Jacsó (2008b), although there was no mention of the fact that the Google Scholar team seemed to have declared unofficially that these numbers were

Table I. Mythical quotes by Péter Jacsó in the column "Savvy Searching" in *Online Information Review*

Article	Quote
Jacsó (2012d)	The parsing and citation matching components require brain surgery
	It is more like a contemporary version of the Aesop fable about the fox who invited the stork for a dinner, and served soup in a very shallow dish
Jacsó (2012c)	Its secretiveness [GS developers] about every aspect of Google Scholar is on par with that of the North Korean government
	One can almost see the scene (and hear the song) from "The Wall" as the grossly under-educated crawlers and parsers march to their destination sites, proudly singing "we don't need no education", and thinking "we don't need no metadata"
Jacsó (2012b)	Its many serious shortcomings make it as inappropriate for scientometric purposes, as making King George III, a.k.a. the "mad king" of Great Britain, the first ruler of the United Kingdom
	GS has acted since its launch as the ultimate mismatchmaker
	GS, the emperor of Secretiveness
	GS—in spite of its name—is omnivorous
Jacsó (2012a)	Using its reported publication and citation counts to calculate metrics, is like inviting a street card shark to be the dealer in a 5-star casino
	I am graced, especially as I am also listed as Albert Einstein's co-author
Jacsó (2010)	The GS parsers are very unconventional but versatile in interpreting any numeric data as a publication year
	Cleaning it up would require much more than spraying out some deodorant and replacing the carpet messed up by the parser puppies of GS again and again. It needs a complete fumigation in the kennel and the GS mansion
Jacsó (2008c)	I have not seen any professional information service that would behave in such a senseless way
	Its pathetic software has a long way to go to make use, at a scholarly level
Jacsó (2008b)	In some European countries omitting the author name from the publication is infringement of the moral component of copyright, an unknown concept in US copyright law
Jacsó (2006a)	In GS searching by journal name is a Sisyphean task
	These errors of artificial unintelligence in matching cited and citing references one hopes will be noted by the natural intelligence of real scholars and practitioners
	That domain [beyond 1,000 hits] remains the world of 1,001+ nights, with equally loose number of promised virgins and harem dancers
Jacsó (2005c)	Unsystematic, unpredictable and disturbingly fragmentary coverage
Jacsó (2005a)	G-S is a free service, and for many who consider it to be a gift for the world it may be anathema to say any but good words of it
	G-S gives a bad name to autonomous citation indexing. It shows lack of competence, and understanding of basic issues of citation indexing

not correct (a fact that was mentioned by Notess, 2008). In any case, it seems that there was some decline in the initial euphoria amongst various experts about the potential of Google Scholar. A notable example was Dean Giustini, who had started a blog dedicated to Google Scholar⁵, and who admitted that "Scholar is not as useful as promised" (cited by Jacsó, 2008b), in reference to the inability of Google Scholar to resolve the limitations that had existed since its launch in 2004. Giustini went on to state that "unless it changes its course, GS will go the way of the dodo bird eventually". Google Scholar did indeed change.

The evolution of Google Scholar was slow, especially during its first years of existence. This may have been due to the fact that the team at the beginning was made up of only two people (Orduna-Malea et al., 2016). In fact, some of the limitations or criticisms that it received in its early days, such as coverage and speed of indexation (Jacsó, 2005a), were later transformed into strengths (Moed et al., 2016; Thelwall and Kousha, 2017, in press).

In 2008, some of the Google Scholar errors that had led to erroneous results and citations, or large-scale duplication thereof, were corrected, "which is

the appropriate reaction to the criticism" (Jacsó, 2008c). A number of other errors could also no longer be reproduced, although many others of a similar magnitude still remained after an apparent cleaning-up of the data. This seemed to indicate that when bad practice was exposed in the press, Google Scholar fixed it so that users could no longer find the exact examples that were reported; they therefore tended to think that the issues had been resolved, although this was not entirely true (Jacsó, 2010). Not only did they persist, but all Google Scholar-based evaluations that had been conducted previously would have irreparably harmed both the individuals and journals that were evaluated.

Jacsó (2010) also complained of the lack of gratitude shown by the Google Scholar team for his and other authors' contributions to correcting the errors, something that had in fact occurred in the case of the Google Books team, which publicly thanked Nunberg (2009) for contributing to the improvement of that tool with his criticism. Another significant complaint was related to Google Scholar's tendency to blame its errors on publisher metadata rather than its own parser, similar to the Google Books team's excuses for errors as reported by Nunberg, as described by Oder (2009), who reproduced the letter from Google in response to the query about the errors detected: "Without good metadata, effective search is impossible".

However, Google Scholar continued to evolve and to grow until, on its fifth anniversary (2009) it eliminated the "beta" tag that it had retained since its launch (Jacsó, 2010), and many of its systematic errors were fixed (corrected or deleted). Subsequently, Jacsó (2011) reported that the Google Scholar parser had improved, such that tests carried out in mid-November 2010 did not detect some of the previous errors, and many others were reduced significantly, although he did continue to warn that it was not yet reliable enough to be used to calculate bibliometric indicators in the evaluation of research activity. Finally, Jacsó (2012a) recognised that the volume of errors was insignificant when compared to the errors identified at the beginning, although the affected authors would not be of the same opinion. A few years later, in Jacsó's prologue to *La revolución Google Scholar: la caja de Pandora académica* (Orduna-Malea et al., 2016), he contended that the reduction in the number of errors, even when positive, was manifestly insufficient since errors persist due to functional issues with the system that have not been resolved.

Indeed, despite the fact that 2011/2012 was a milestone in the history of Google Scholar with the emergence of the related services Google Scholar Citations (aimed at authors) and Google Scholar Metrics (aimed at journals), and definitive growth

in its coverage and speed of indexation, many of the errors reported during the 2004 to 2012 period still persist today.

Rationale and objectives

Given the growing use of Google Scholar not only as a gateway to searching for academic literature, but as a bibliometric tool, the identification, classification and quantification of its errors and limitations when calculating bibliometric indicators is of paramount importance.

However, scholarly literature dedicated to this matter has not been systematic. With the exception of Jacsó, few authors have directly sought to detect, describe or gauge the influence of errors in Google Scholar. Occasionally, these limitations have been given passing mention in certain publications, but they have received scant attention in the way of description or explanation or have been quite simply overlooked.

Moreover, in many cases these limitations have been mistaken for errors, when they are in fact related but different aspects. The limitations of Google Scholar are related to certain services or features that prevent it being used as a bibliometric analysis tool. These limitations include not being able to sort the results by the number of citations or the year of publication, the absence of an API (Application Programming Interface), the maximum of 1,000 results per search or limited capabilities for exporting search results, to give only a few illustrative examples. The objective and purpose of Google Scholar is not bibliometric analysis but searching for scholarly literature. Therefore, if such analysis is tedious, we should mark it down as a mere limitation but not as an error.

Conversely, an error arises in relation to features that Google Scholar should provide or execute correctly if it is to fulfil the goals and tasks that it officially declares itself to offer. For example, the system claims to report the number of citations received by a publication from the other publications indexed on Google Scholar. Therefore, if this number is incorrect, we have located an error. Since these functional errors directly affect the calculation of bibliometric indicators, knowing what types of errors exist, and to what extent, are important challenges in present-day bibliometrics.

This study is therefore a first step along this line of research. Its main objective is to carry out an exhaustive bibliographic review of what has been said and done about errors in Google Scholar, to then categorise the findings of the studies that we have included in our review.

2. METHOD

The bibliographic review of errors in Google Scholar was conducted over three consecutive phases. First, empirical studies on Google Scholar were compiled. Second, the studies that addressed errors in Google Scholar, either directly (as part of the objectives) or indirectly (errors were listed or described even if they were not part of the main objectives), were selected. Finally, the selected studies were qualitatively analysed in order to group them according to error type.

The first phase (compilation of empirical studies) was carried out as part of the objectives of a nationally funded research project (HAR2011-30383-C02-02). For this purpose, an online information and bibliographic review service was created, called Google Scholar's Digest (<http://googlescholar Digest.blogspot.com.es>), which has been compiling all empirical studies that provide data on Google Scholar since 2014, offering critical reviews (digests) of the most relevant studies.

This service was put together from systematic searches of the main bibliographic databases (WoS, Scopus and Google Scholar itself) and is constantly fed by a technological monitoring and alerts system, using RSS technology, a Twitter account (@GSDigest), and the Google Scholar alerts system. To date, 271 publications have been compiled, including journal articles, books, book chapters, conference papers, reports and working papers, among other document types.

This system was designed in part because of the complexity of finding academic literature with empirical data on Google Scholar, since searches limited to the term <Google Scholar> in the title, keywords or abstract are insufficient.

The second phase (selecting the studies on errors in Google Scholar) consisted of a qualitative analysis of the 271 publications in the Google Scholar's Digest bibliography. The studies were separated into two distinct corpuses. On the one hand, the work of Péter Jacsó (Corpus A, comprising 16 works, see Appendix I) and, on the other hand, other studies with data or comments on errors in the functioning of Google Scholar (Corpus B, comprising 34 works, see Appendix II).

The third phase (categorisation of errors) consisted in the reading, analysis and manual classification of each of the studies in the two bibliographic corpuses, in order to identify both the different currents in the literature on errors and the main types of errors studied to date.

To this end, we decided to apply a general taxonomy of errors (Table II), in order to classify the studies according to the type of error addressed (note: a study may, of course, contain information on several types of errors).

Table II. Broad taxonomy of errors in Google Scholar database

Error type	Scope
Coverage	Related to the erroneous inclusion/exclusion of material in the database based on accepted and rejected typologies
Parsing	Related to the misidentification of the descriptive elements of a bibliographic record
Matching	Related to the linkage between bibliographic records
Searching & Browsing	Related to the search features of the search engine

Phase I was carried out from 2014 to May 2017, while phases II and III were carried out in parallel between January and May 2017.

3. RESULTS

This section is divided into four main blocks. First, a descriptive analysis of the bibliographic corpus is carried out. Second, studies focusing on the identification and description of errors in Google Scholar are examined. Third, publications that have focused their interest on errors in filtered or structured environments – either official services (Google Scholar Citations, Google Scholar Metrics) or existing tools in the market (Publish or Perish) – are looked at. Finally, the publications that have proposed Google Scholar error type categories are singled out.

3.1. Descriptive analysis of the bibliographic corpus

As mentioned previously, the literature that has dealt with errors in Google Scholar was divided into two bibliographic corpuses. The first (corpus A) comprising the work of Jacsó (16 publications, Appendix I), and the second (corpus B) comprising other publications that have addressed, directly or indirectly, the issue of errors in this database (33 publications, Appendix II), forming in total a corpus of 49 publications.

Of the total number of publications, 40% (20) are concentrated in the period 2005 to 2008, corresponding to the launching of the search engine and the bulk of the articles published by Jacsó, who after then authored an annual review for his column in *Online Information Review* (2009-2011). 2012 is an exception (four works by Jacsó), coinciding both with the update of the search engine and the birth of Google Scholar Citations and Google Scholar Metrics. From then on, Péter Jacsó ceased his fertile output dedicated to Google Scholar. Corpus B, for its part,

developed strongly during the first years, although no remarkable pattern is observed. One possible reason for this is that a significant proportion of these publications did not focus on the errors of Google Scholar, which nevertheless appeared during their development; the errors were then only reviewed in passing (in varying levels of detail). In any case, the number of publications grew in 2016 (five in total).

With regard to thematic coverage, 53% of all the publications (corpus A and B) focus on specific disciplines while the remaining 47% are multidisciplinary studies. These data are influenced by Jacsó's work, as 12 of his 16 studies (75%) cannot be ascribed to any disciplinary area, since they are based on the testing of different search options through general queries. As far as geographic coverage is concerned, 76% (37) of the publications are international in scope, while only 24% (12) focus on specific countries. Again,

Jacsó's work influences this distribution since all his articles have an international approach (or rather, they have no geographical restrictions). Finally, most of the publications have analysed authors (41% of the total), followed by journals (25%) and documents (17%). Figure 1 gives a summary of the descriptive data of the analysed bibliographic corpus.

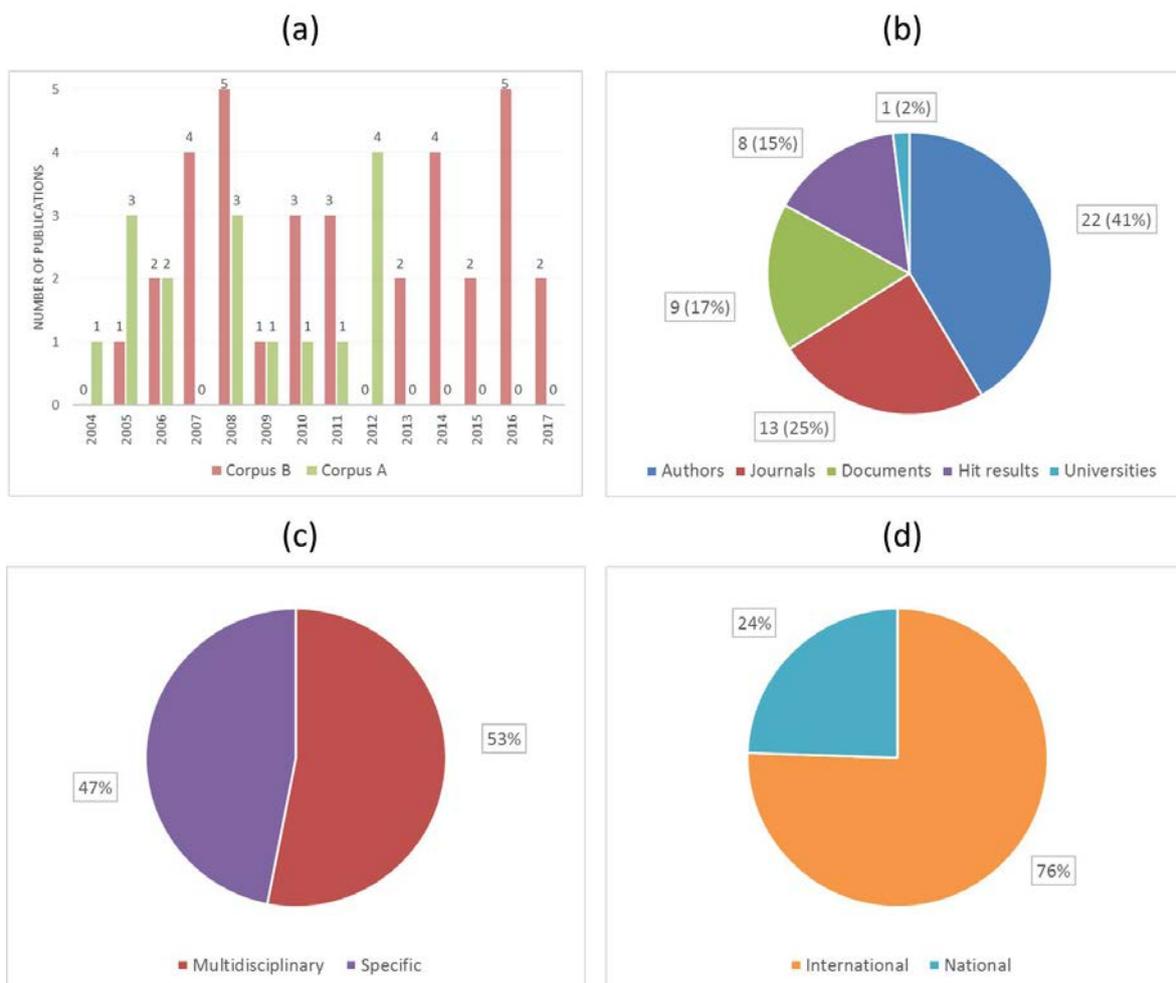
3.2. Errors in Google Scholar

Following the scheme proposed in Table II, contributions were classified into those that identify errors related to coverage, parsing, matching and searching.

3.2.1. Coverage

Given the scarce – at times, inexistent – information on the sources that feed Google Scholar

Figure 1. Descriptive analysis of the bibliographic corpus on Google Scholar errors (a) annual output; (b) unit of analysis; (c) thematic coverage; (d) geographical coverage



(Jacsó, 2012c; Orduna-Malea et al., 2016), critical literature on its coverage was particularly fertile during the early years of its existence. Jacsó (2005a), from the outset, reproached it for the fact that the results for any query were made up of a mixture of document genres (journal paper, conference paper, or book) and paper types (research paper, review paper, brief communication) from a multitude of sources, including not only educational websites but also non-scholarly sources, like promotional pages, table of contents pages, course reading lists (Jacsó, 2006a).

The academic literature has sometimes treated this as an error when in fact we are faced with a limitation for conducting certain bibliometric analyses. Moreover, it is not even globally accepted as such because many specialists consider that the varied nature of the citing documents is not necessarily a limitation in itself.

However, having performed several tests to verify the validity of the system with such an amalgam of citing documents, several errors related to coverage were discovered incidentally:

- Massive content omissions when searching for journals (Jacsó, 2005c): this even occurred with the journals of publishing houses that had agreements in place with Google Scholar to display the full text of the contents (*Nature*, *Science*, *PNAS*).
- Indexing limits (Price, 2004; Jacsó, 2005c): limits were detected in the indexing of files, set at the first 100 to 120 KB of the text, such that, if the terms of a query appeared in the text beyond that limit, a result might not be returned and, therefore, the corresponding hit would not be counted.
- Mistaken inclusion of excluded document types (Jacsó, 2008a): sometimes a book review was mistaken for the book itself. Apart from the errors made due to wrongly classified document types, the coverage policy of Google Scholar was also brought under scrutiny: "Content such as news or magazine articles, book reviews, and editorials is not appropriate for Google Scholar"⁶.
- Inclusion of excluded document types due to mass indexing (Jacsó, 2012c): when Google Scholar considered a web domain for inclusion (for example, .edu), it indiscriminately indexed all the files hosted in that web domain that were apparently academic, which led to the indexing of many document types that in principle, according to its rules and criteria, were not appropriate for the database.

3.2.2. Parsing

Parsing errors are one of the most important areas of this study, as their occurrence causes a chain reaction that is capable of generating and transmitting new errors to other documents on an extremely large scale. Parsing is a process that enables strings of symbols to be analysed according to predetermined formal grammatical rules. Hence an application can identify the different parts of a bibliographic record (author, title, source, volume, number, pagination) of both a citing document (metadata) and a cited document (bibliographic reference contained in the bibliography section of an academic work).

Belew (2005) had already indicated that certain character encodings, such as ASCII, can generate problems and errors (inconsistencies in author names and erroneous attribution of citations) in WoS and Google Scholar, especially for authors whose names are written in non-Latin characters. However, Bar-Ilan (2006) expressed surprise when, in performing a bibliometric analysis of the scholarly output of mathematician Michael Rabin, she discovered that there were recurring errors (erroneous attribution of citations and authors) in articles published by the IEEE (Institute of Electrical and Electronic Engineers), even though Google Scholar – supposedly – based much of its data on the information provided by publishers. In reality, the main problem with Google Scholar was related to the fact that it programmed its own parsers instead of relying on the information provided by the metadata prepared by publishers, an approach that may make sense with unstructured masses of web pages, but not with scholarly documents (Jacsó, 2005b; Jacsó 2012c), leading it to generate enormous amounts of errors during the process of scanning and parsing the various elements of a bibliographic record. This fact led to the discovery that the author "I Introduction" was the most prolific according to Google Scholar, with more than 40,000 publications (Jacsó, 2006a) or that "F Password" was the most cited (Jacsó, 2008b). The faulty functioning of the parsers led to segments of the International Standard Serial Number (ISSN) being mistaken for the year of publication (Jacsó, 2008a), and menu options, section headings and journal name logos for author names (Jacsó, 2009a), due to the complete lack of quality controls (Jacsó, 2010), distorting the bibliometric indicators at individual, corporate and journal levels (Jacsó, 2012c).

As an illustrative example, and in response to Vanclay (2012), Jacsó (2012b) showed a result obtained by Google Scholar for the article "Vision 2020—the palm oil phenomenon", in which the

system deleted the second author (MA Simeh), showed "Growth" as the publication source (when in fact it was the *Oil Palm Industry Economic Journal*), and "2015" as year of publication (when it was actually 2005). Figure 2 shows the current result for this article with its corrected bibliographical data.

Within the parsing errors, the literature has dealt with each of the elements of a bibliographic record, although errors related to author names have undoubtedly been the most widely studied. For that reason, we shall now look at author studies separately from the other bibliographic elements.

a) Absurd authors

Péter Jacsó (2004) denounced the irregular and deficient behaviour of the Google Scholar parsers from the outset, especially when identifying author names, which were confused with other content (Jacsó, 2008a). Marydee Ojala (2005) expressed similar sentiments in a brief text included in the article by Wleklinski (2005), published in the journal *Online*. Harzing and Van der Wal (2008) also contended that Google Scholar would not find publications if the author's name included a sequence of characters that was not in a traditional typeset or if the author had used LaTeX (a document preparation system).

On occasion, a "misspelled author" error was generated, whereby names such as "Julie M Still" became "Julie M" or Péter Jacsó himself became "Peter J", such that the first letter of the surname became the first initial of the forename (Jacsó, 2008b).

On other occasions, nonexistent names were generated. Jacsó managed to identify a large number of these, such as: Payment Options, Please Login, Strategic Plan, I Background and II Objectives, Forgot Password, I Introduction and R Subscribe, among many others (Jacsó, 2008b; 2008c; 2010; 2011). These errors were sometimes concentrated in the publications of certain publishers, such as

Emerald (Jacsó 2008b), or journals such as *The Lancet* (Jacsó, 2010), where parsers sometimes created author names from the MeSH terms (Medical Subject Headings) assigned to the documents. Even though Jacsó (2010; 2011) acknowledged that in some cases these names may be legitimate (notably the case of Raymond and Linda Measures), most of the time they were large-scale errors: V. Cart corresponded on most occasions to View Cart, and not to Veronica Cart (Jacsó, 2008c).

Table III provides a comparison of the results obtained (number of hits returned) for a query by absurd author (example: <author:"F Password">) in Google Scholar in the different publications that have addressed the subject, including the results obtained in 2017 for the purposes of this study.

As has already been mentioned, sometimes these terms were real (Jenice L View) and other times they were parsing errors, which substitute (VIEW, TPO, from VIEW, TIONAL POINT OF) or modify (KALINGA, AVF, from KALINGA, A View From) or add (Image, PVVS, from Physically-Valid View Synthesis by Image). These absurd authors still exist as of 2017.

Finally, on other occasions co-authors (real or absurd) were added. Jacsó (2008b) denounced the fact that in the bibliographic record corresponding to the seminal article on h-index published by Jorge Hirsch (2005), Google Scholar had added three co-authors (Louie, Jackiw and Wilczek), who were the researchers that Hirsch used as examples in his article within an enumerative list (this result has been updated and is now correct). Jacsó himself also fell afoul of this quirk in the search results, appearing in the company of "MA Sicilia" as co-author of his article "Deflated, inflated and phantom citation counts" (Jacsó, 2006a). Curiously, this erroneous information only appeared in what was considered to be the main version, but was correctly recorded in the other versions (Figure 3).

Figure 2. Correction of the bibliographic description of results in Google Scholar



Table III. Test search on absurd authors

Query	Jacsó (2006a)	Bar-Ilan (2008)	Jacsó (2008b)	Jacsó (2008c)	Harzing & Wal (2008)	Jacsó (2009a)	Torres et al. (2009)	Jacsó (2010)	Jacsó (2011)	Jacsó (2012c)	Replicated hits (2017)
Access										78,100	6,840
Assessment									45,900		16,100
Contents										25,200	11
Design									74,000		59,800
Evaluation									43,900		2,720
F Password			102,000								0
I Introduction	40,100	689,000	*5,990		956						5
Introduction						42,400				40,100	148
M Profile			12,400								137
Methods									42,300	42,700	1,030
Password						910,000			910,000	910,000	1
Policy								120,000	105,000		56,800
Subscribe									73,400	77,600	155
Therapy									49,500		5,420
Treatment								39,500			7,770
V Cart				85,000		85,300					16
View										1,100,000	3,750
Introduccion							7,940				23
Estado de la cuestion							1,330				12
contenido							6,180				416
indice							4,320				1,813
i-netlibrary							12,200				0

* Note: the unusually low value compared to the rest may be due to the query performed (using the <author> command or not)

Figure 3. The addition of a phantom co-author to a bibliographic record in Google Scholar



b) Other bibliographic fields

Within this area, we may highlight the publications that reflect errors in titles and bibliographic information (mainly, name of journal, volume, number and pagination):

- With regard to the document title, Jacsó (2005b) contended that it was sometimes mistaken for sections or subtitles ("Short Communication", "Original Article" or "Special Invitation"). These elements could be added to the original title or replace it completely. The reason for this error lay in the fact that Google Scholar ignored the metadata and focused on detecting sequences of characters with some special emphasis (boldface, larger font size, etc.). Walters (2007) is one of the few authors that has given a figure to this type of error, after evaluating 155 articles and detecting that 15.5% (24 documents) of them had incomplete titles. However, it would seem that he used the snippet provided under each result to perform this analysis. Walters then stated that GS did not include more than 4 authors and no more than 99 characters (in reality, GS uses a line to include these data, and the author can now access the complete reference by clicking on the "cite" option), so we may assume that the error rate that he obtained was an overestimation. For this reason the title is sometimes shortened, as shown by Bar-Ilan (2008) in her analysis of the publications of the American Physical Society, in which she

identified inconsistencies in the publication dates.

- With regard to the publication date, Jacsó (2008b) discovered errors because the parsers identified any chain of 4 digits as a potential publication date, including page numbers or area codes or street addresses in author affiliation. For example, the volume number was sometimes used to designate the publication date, as Jacsó (2010) pointed out in the case of the "Proceedings of SPIE", or the year of the latest edition of a book was mistaken for its publication date (Dilger and Müller, 2013; Martín-Martín et al., 2017). On other occasions, there was simply no date of publication (Jacsó, 2010), a fact that was also reported by Maia et al. (2016), who, after analysing 2,400 documents in the area of "Strategy as Practice", noted that 15% of the documents had no publication date on Google Scholar. These errors led the system to absurd situations in which some documents had future publication dates (Jacsó, 2008c), which in turn caused "future" documents to have already been cited by other documents. Nevertheless, these publications are small samples that were compiled in the Google Scholar's developing years. In a later study (Martín-Martín et al., 2017) of a sample of 32,680 highly-cited documents, the authors confirmed that there is agreement between the publication dates reported by Google Scholar and Web of Science for 96.7% of documents. Although WoS is not error-free,

the fact that it is a supervised database gives us certain guarantees regarding the quality of Google Scholar data today.

- The document source is another field that has been explored by the literature. Jacsó (2005b) had already indicated that there were results that did not provide information about the source, even when they originated from Medline. Subsequently, Maia et al. (2016), who worked with a sample of 633 records, indicated that 27% of them contained no mention of their source.
- These parsing errors do not only affect source or citing documents, but also documents cited by them. Meho and Yang (2007) analysed citations received by 25 professors from Indiana University-Bloomington, demonstrating that 475 citations from Google Scholar did not have complete bibliographic information, although it should be noted that these citations came from unusual document types (presentations, grant and research proposals, doctoral qualifying examinations, submitted manuscripts, syllabi, term papers, working papers, Web documents, preprints, and student portfolios). Similarly, Noll (2008) studied the coverage of Google Scholar in the area of art history literature, highlighting the existence of errors in the cited references, which lacked information on the volume, number and pagination.

The reasons for which the Google Scholar parsers commit these flagrant errors have been very little studied, beyond the work of Jacsó. One such study that merits our attention is that published by Haddaway et al. (2015), who, after investigating the usefulness of Google Scholar as a database in systematic reviews and grey literature, calculated a total rate of duplicate records due to parsing errors of around 5%, because of the following factors:

- Typographical errors introduced by manual transcription (15% of title records).
- Differences in formatting and punctuation (18% of duplicates).
- Capitalisation (36% of duplicates).
- Incomplete titles (15% of duplicates).
- Automated text detection (3% of duplicates).
- Scanning of citations within references of selected included literature, and the presence of both citations and the articles themselves (13% of duplication).

3.2.3. Matching

In most cases, matching errors are derived from parsing errors, since small variations in a reference can lead to duplicate records (Harzing and Alakangas, 2016), although they are sometimes errors in themselves. In any case, the consequences of these errors for bibliometric analysis are enormous, especially because of the fact that they generate a mass of inflated document citations. As an illustrative example, Jacsó (2008b) analysed his own article "Google Scholar: the pros and the cons", which at that time had received 57 citations according to the Google Scholar result. However, after exhaustive filtering of the data, Jacsó found that this figure was highly inflated. First, the number of estimated hits was 55, of which the interface actually displayed 53 (this is occasionally due to desynchronisation caused by a database update). Of these, there was no way to access four of them (their veracity could not therefore be verified), six were duplicates and four others were erroneous (citing document did not mention the cited document).

This example alone would suggest to the reader that there is a wide variety of interconnected errors, both in matching and browsing (see next section). Although the errors should be studied in terms of their cause-effect relationships, the literature has generally treated them separately, distinguishing between matching errors between different versions, on the one hand, and matching errors between citing and cited documents, on the other.

a) Matching versions

Duplicate versions of records are an issue that have been brought to light by the literature practically since the launch of Google Scholar. Jacsó (2005b) illustrated the existence of different versions of the same document that were not correctly linked and how this caused dispersion in the citations received by a document, which ultimately affected the position in which that document appeared in the results.⁷ Yang and Meho (2006) also commented on how a citation from two versions of the same document (preprint and the version of the article published in a journal) would be counted twice. However, studies that have provided exact figures that quantify the magnitude of these errors in a particular sample or in Google Scholar in general are very scarce, and with completely different results due to the enormous differences in the samples used. Noll (2008) detected 23% of duplicates and multiple versions that contributed to the number of citations received by a set of 12 preselected art historians. Rosentreich and Wooliscroft (2009),

after calculating the g-index for a set of 34 accounting journals, detected a duplicate rate of around 3%. Thor and Bornmann (2011) described how, in the case of a specific search (<allintitle: merge purge large>), they obtained eight results in Google Scholar, and all referred to the exact same document, which ironically dealt with the automatic identification of duplicates.

However, it should be noted that the system for automatically identifying versions has improved substantially over time, an aspect to which Google has dedicated technological resources, as can be seen through the publication of a patent that describes the automatic identification of different versions of the same document (Verstak and Acharya, 2013).

The article by Pitol and De Groot (2014) was the first dedicated exclusively to the issue of versions in Google Scholar. The authors analysed 982 articles, concluding that only 6.1% of them (60) had duplicate versions, which was taken to mean that they were documents that the system had not merged. Moed et al. (2016) also indicated that duplicates, in the strict sense (with identical metadata), were rare (0.2%) in their study of a limited set of articles (1200) published in 12 journals. Even so, this percentage depends on the document type analysed, increasing significantly in the case of monographs. Martín-Martín et al. (2017) analysed the article "Mathematical Theory of Communication", for which they detected up to 165 versions that were not correctly linked.

b) Matching citing/cited documents

Another source of error is the matching of citing (source) and cited (target) documents. Although citations are prone to many forms of error (e.g. typographical errors in the source document because authors or journal editors have incorrectly transcribed a bibliographic reference), other problems are caused by the Google Scholar parsing process, especially when non-standard reference formats are used (Harzing and Van der Wal, 2008) or when the document has a complex structure (Meho and Yang, 2007), or simply when the parsing process fails. In the words of Vaughan and Shaw (2008), "citing and cited papers are confused".

The Google Scholar automatic citation system functions correctly when a bibliographic reference exactly matches a master record (Jacsó, 2009a). In that case, it is rewarded with a new received citation. However, it may be that there is no such match because the parsing has generated variants or duplicates, both of the reference and the master record (or both). If the version-linking technology

(mentioned above) worked correctly, many of the errors would be resolved, although this regrettably is not the case.

Jacsó (2005b) was the first to write about the notorious inability of Google Scholar to correctly link citing/cited documents, resulting in an inflation/deflation effect in the cited documents (Jacsó, 2008a), due to either receiving citations that do not exist or not receiving existing citations. For example, Jacsó noted that the most-cited article in *The Scientist* was a document with 7,390 citations received which, in reality, corresponded in large measure to an article published in the *Journal of Crystallography*. Subsequently, Harzing and Van der Wal (2008) were not able to reproduce this search, and they noted that the most-cited article was another (which received only 137 citations), from which it follows that Google Scholar was able to correct this error.

In spite of this, the reporting of errors in empirical studies is notable. Meho and Yang (2007) observed that Google Scholar missed 40.4% of the citations listed in both WoS and Scopus for 25 professors, and Bar-Ilan (2008) noted that the article "Probabilistic Encryption", cited 915 times, had been attributed incorrectly to Avi Wigderson. Jacsó (2008b) pointed out that most of the citations received by an article published in the *Journal of Forestry Ecology & Management* actually cited a technical report, yearly updated, that had part of the same title as the journal article. Which meant that "GS lumps together a series of technical reports and a journal article, awarding the citations to the journal" (Jacsó, 2008b).

At other times, the matching error stems from an earlier parsing error. For example, Jacsó (2008b) reported that the authorship of an article published in *Online Information Review* was attributed to "M Profile" when in fact it was co-authored by Hong Iris Xie and Collen Cool. Since this article had received 10 citations, the two authors had been deprived of these citations. If "I Introduction" had been the author of around 6,000 articles in Google Scholar (see Table III), the number of citations that the actual authors did not receive could be in the millions; it is as impossible to calculate as the number of wrongly attributed authors. The direct consequence is that the citation/matching algorithm is as unreliable as the parsing algorithm. These errors, even if they have been minimised, still exist. For example, Moed et al. (2016) indicated that one of the most-cited articles in the *Journal of Virology*, according to Google Scholar (270 citations), received most of these citations (180) erroneously.

3.2.4. Searching & Browsing

The last aspect that remains to be described is general errors associated with the search and browsing processes in the Google Scholar environment. This type of error has sometimes been confused with or placed alongside search limitations. In this case, we shall only highlight those contributions that look specifically at errors.

From the qualitative analysis of the bibliographic corpuses of errors in Google Scholar, we separated out the contributions that report errors in the advanced search due to a lack of authority control, in the number of hits for a query, and errors in the full-text link.

a) Advanced search

As might be expected, the pioneer in this field was Jacsó (2005a). When he conducted a bibliometric analysis of Garfield's work to coincide with his 80th birthday, he discovered a series of deficiencies due mainly to the absolute lack of authority control (Bar-Ilan, 2008), which generated errors in searches by author (the system combined the publications of E Garfield and RE Garfield, for example) and by journal (the system combined all articles published in *Current Science* with those of other publications in which the same character string appeared, such as "Current Directions in Psychological Science" or "Current Trends in Theoretical Computer Science") (Jacsó, 2005a). This is an error in the sense that the database was unable to return the articles published by a particular author or journal, which is the service that had been promised to the user. At present, at least for *Current Science*, this error seems to have

been resolved, although authority control is still lacking (a search for "revista española" ("Spanish journal") will retrieve articles published by *Revista Española de Lingüística Aplicada*, *Revista Española de Pedagogía*, *Revista Española de Documentación Científica*, etc.) and is complicated by the existence of abbreviations and variants (Jacsó, 2006b), a problem that still occurs.

In its beginnings, Google Scholar provided an advanced search function to look for documents according to their discipline. Jacsó (2008a) revealed this to be an absurd function, since a search not restricted by subject generated 85% more results than adding up the results for each of the categories.

b) Hit estimate errors

Within the errors in hit estimates, the literature has mainly dealt with errors based on queries using Boolean logic, the duplication of hits, and advanced search publication date.

Boolean problems

This type of problem was a classic example in Jacsó's work. They are problems related to absurd or inconsistent numbers of results according to the query. For example, the search for "protein" returned 8,390,000 results, the search "proteins" 4,270,000, and finally the search for "protein OR proteins" 1,630,000 (Jacsó, 2005a). Based on this study, we have compiled all the examples provided throughout the work of Péter Jacsó and recalculated these data for the present day (Table IV). In this way, we may see how the errors not only persist but, in some cases, have even increased.

Table IV. Test search on absurd Boolean hit counts

Work	Query	Original hits	Replicated hits (2017)
Jacsó (2008c)	Lancaster	442,000	1,130,000
	Lancaster OR Lancaster	348,000	170,000
Jacsó (2006b)	chicken breed	9,750	95,800
	chicken breeds	5,300	59,600
	chicken breed OR chicken breeds	5,300	141,000
	Chicken	387,000	2,300,000
	Chickens	168,000	1,070,000
	chicken OR chickens	322,000	2,170,000
Jacsó (2006b)	Scholar	556,000	6,520,000
	Scholars	939,000	2,510,000
	scholar OR scholars	611,000	4,720,000
Jacsó (2005a)	Protein	8,390,000	5,270,000
	Proteins	4,270,000	4,550,000
	protein OR proteins	1,630,000	5,190,000

Duplicate hits

The generation of repeated hits has also been a recurrent issue in the Google Scholar literature (Jacsó, 2005a; 2006b; 2008b; Shultz, 2007): the existence of duplicate records in Google Scholar results due to parsing and matching errors (versions). It should be mentioned, however, that much of the literature uses erroneous terms when referring to the concept "hit" (results for a specific search), because sometimes they use it as a synonym for "citation" (citations aggregated under a master record), although they are related but different concepts (Levine-Clark and Gil, 2009). It is therefore difficult at times to follow or appropriately contextualise many of the findings and conclusions. Of the few publications in which specific figures are given, Jacsó (2008b) reported how, after analysing the number of articles published in *Online Information Review* indexed by Google Scholar, he obtained a total of 513 records (thus hits). Of these, approximately 38% (195) were duplicates, with the added problem that this figure (513) varied depending on the Search Engine Result Page (SERP) that the user was on at any given moment.

Year range

If parsing errors are assumed in the publication dates, we could not expect an advanced search by publication date to be error-free. Table V compiles all the examples provided by Jacsó with a reconstruction of the searches for 2017 conducted for this bibliographic review of errors. As can be seen, inconsistencies still persist.

c) Erroneous full text links

Finally, the literature has identified errors in the links in the master records that provide access to the full text of the article, where this is possible. Jacsó (2005a) found that by clicking on the link to an article published in 2005 on *Infection and Immunity*, the system took him to the full text of another article published 25 years before in *PNAS*. Likewise, Shultz (2007) discovered the existence of broken links or dysfunctional links. Later, Martín-Martín et al. (2016b), after conducting an exhaustive case study into the article "Mathematical theory of communication" on Google Scholar, detected 830 versions linked to the master record. Of these,

Table V. Test search on absurd time range hit counts

Work	Query	Time range	Original hits	Replicated hits (2017)
Orduna-Malea et al. (2015)	<empty query>	1700-2013	596,000	699,000
		1750-2013	567,000	717,000
		1800-2013	552,000	758,000
		1850-2013	566,000	766,000
		1900-2013	541,000	*676,000
		1950-2013	617,000	742,000
		2000-2013	693,000	739,000
Torres-Salinas et al. (2009)	<empty query>	2006-2008	93,900	915,000
		2005-2008	89,800	813,000
		2004-2008	139,000	*930,000
		2000-2008	109,000	671,000
Jacsó (2008c)	Lancaster	1907-2007	116,000	386,000
		1917-2007	118,000	*350,000
Jacsó (2008a)	Vietnam	1435-2008	20,200	993,000
		1960-2008	20,600	*911,000
Jacsó (2006a)	Scholars	2005	23,600	188,000
		2005-2006	11,900	181,000
Jacsó (2005a)	<empty query>	1995-2005	976,000	738,000
		1985-2005	966,000	672,000

* Note: these results are coherent with Boolean algebra although quite implausible

Google Scholar only returned 763, of which 21.1% (161) presented some kind of error. In particular, 86 had a broken link to the full text.

3.2.5. Global error propagation

The errors identified by the scholarly literature analysed in this study have barely been quantified, and most of the time they are merely mentioned or reported. Despite the absence of error percentages, the deficiencies were sufficiently voluminous for Jacsó (2008c) to conclude that the citations reported by Google Scholar were not acceptable, not even as a starting point, for the evaluation of the scholarly activity of researchers, since the volume of citations was "inflated" and "untraceable", which had similar repercussions for the calculation of derived indicators such as the h-index (Jacsó, 2009a; 2012c).

To illustrate these shortcomings, the literature has carried out several analyses that have revealed the combined occurrence of several types of errors that distort the overall results, among which the following publications stand out:

- Jacsó (2009a): analysed the book *Managing the Multinationals: an International Study of Control Mechanisms* (Harzing). Seven (unlinked) versions were detected, each with its own received citations (citation dispersion), and one result corresponded to a book review that was erroneously attributed to Harzing.
- Bar-Ilan (2010): analysed the book *Introduction to Informetrics* (Egghe and Rousseau). 358 documents referring to the article were detected, of which 24 (6.7%) were duplicates, 17 contained title variants, and 5 had authorship errors. After removing the duplicates and other errors, only 307 documents actually cited the book (total error of 14.2%).
- García-Pérez (2010): analysed a corpus of 380 publications by 4 authors in the field of psychology. 16.5% of the citations were erroneous (phantom citations, duplicate links, unlinked versions and errors in the estimation of hits).
- Adriaanse and Rensleigh (2011): analysed the content of 9 environmental science journals in South Africa, identifying a total of 448 inconsistencies in the records (14%) as well as duplications (a total of 185) due to "citation" hits in Google Scholar. Note: In this case, hits are counted as citations (citation hits).

- Martín-Martín et al. (2014b): analysed a corpus of 64,000 highly-cited documents between 1950 and 2013. The following errors were identified: full-text links that did not work or did not correspond to the master record, GS-WoS linking failures,⁸ unlinked versions, incorrect attribution of citations to documents, incorrect attribution of documents to authors, phantom citations, phantom authors, incorrect identification of titles, duplicate citations and publications.

As can be seen, the broader the samples, the greater the quantity and variety of errors found. This is due, as already mentioned, to the interconnection between different types of errors: a parsing error can generate a duplicate which, if the version control system does not group them correctly, can generate a duplicate citation.

3.3. Errors in filtered environments

All of the studies reviewed above operate in the Google Scholar environment. However, there are platforms, both external and linked to this service, designed for working with more filtered and structured data, which may in some cases help to fix some of the errors seen in the previous sections, although they may similarly introduce new errors.

a) External products

One of the more notable external products is Publish or Perish (PoP) (Harzing, 2010), a desktop application that provides a user-friendly interface for searching Google Scholar directly and, especially, for working with results in such a way that allows users not only to work with the retrieved documents (sort them according to various criteria, merge duplicates, etc.) but also to obtain a wide variety of bibliometric indicators calculated from the retrieved documents. This application, which is free to download and use,⁹ has undoubtedly contributed to the democratisation and popularisation of bibliometrics.

Jacsó (2009a) analysed the first versions of the tool, confirming its potential to facilitate the discovery of erroneous information and correct it. However, since the application works directly with Google Scholar results, it inherited certain errors (e.g. typography errors in author names or errors in the title, phantom authors, phantom citations) and limitations (e.g. a maximum of 1,000 results per query) that cannot be directly corrected or resolved. The ability of PoP to export the results obtained to a spreadsheet can mitigate, but not solve, some of the problems.

Baneyx (2008) developed a complement to PoP called CleanPoP, which works with the results provided by PoP to improve their quality. Its capabilities include the automatic detection and merging of duplicate articles and variants of the author name. As a sample, Baneyx analysed 12 French researchers. Focusing on one of them (R. Br), PoP located 3,707 citations that, after using CleanPoP, were reduced to 526, so the author concluded that about 86% of the citations provided by PoP were incorrect.

b) Internal tools

The Google Scholar team, fully aware of the errors and limitations of their database, developed and launched two new services between 2011 and 2012 that directly draw on the Google Scholar database. First, Google Scholar Citations (GSC),¹⁰ and, second, Google Scholar Metrics (GSM),¹¹ oriented towards the management of authors and journals, respectively.

First impressions of Google Scholar Citations (from an errors point of view) were positive. Jacsó (2012a) admitted that this platform “apparently managed to separate – if not all, but most – of the wheat from the chaff”, since a large number of duplicates were identified and corrected. In addition, the fact that it allowed the authors themselves to correct and edit the descriptive metadata of their articles could help in the medium and long term to improve the quality of the data, so the system was seen as promising. However, many inherited errors were still present (some of which the authors themselves could not correct, for instance separating versions of documents that had been incorrectly merged by the system).

Moreover, Google Scholar Citations has its own errors. For example, in the automatic generation of co-authors, Jacsó (2012a) criticised the fact that in his own list there were authors with whom he had not published: “most of them I have not heard of, let alone known or worked with”. At present, this process has improved considerably, although many of the errors are the result of actions, deliberate or not, of the authors themselves, who, through interest, negligence or incompetence may have incorrectly filled in the various personal information fields or edited the description of a document. The number of citations received per document is a value automatically calculated by Google Scholar, in which authors can not intervene. Even so, there are errors in the processes that are performed automatically. For example, Doğan et al. (2016), after analysing the profiles of 10 researchers from the Department of Information Management

at Hacettepe University, estimated that 55% of their contributions (135) had received duplicate citations, representing approximately 12% of the total number of citations received. Martín-Martín et al. (2016c) also detected duplicate documents, incorrectly merged documents and incorrect titles when analysing the GSM profiles for 814 bibliometrics researchers. Subsequently, Orduna-Malea et al. (2017) detected and classified errors in the automatic linking of authors with their institutional affiliations, in the case of the Spanish university system (wrong by normalised names, disambiguation problems, incorrect linking, multiple official academic web domains, errors with complex, multiple and internal affiliations).

With regard to Google Scholar Metrics, impressions were similar. Jacsó (2012d) described the service as a potentially useful and complementary tool for journals, although he also acknowledged that the information provided, while it is an improvement, is only “plastic surgery”, and that “the parsing and citation matching components require brain surgery to qualify GSM for bibliometric purposes at the journal level”.

Apart from the errors inherited from Google Scholar, GSM also has errors of its own making, such as linking articles to the wrong journals. Jacsó (2012d) was surprised that GS occasionally provided correct data but that, subsequently, GSM attributed an article to the wrong journal. These attribution errors consequently caused errors in the attribution of the h5-index of publications.

Also noteworthy are the annual reports from the EC3 Research Group on the release of each new version of GSM (Martín-Martín et al. 2014a; 2016a). These reports have enabled us to explore a wider variety of errors, particularly those related to normalisation problems (unification of journal titles, problems in the linking of documents, and problems in the search and retrieval of publication titles).

3.4. Classification of errors

The last body of publications on Google Scholar errors has tried to categorise and classify existing errors. However, it should be pointed out that these classifications are not only incomplete (not reflecting all types of errors), but were carried out in a way that complemented or supplemented the original work, the main objectives of which were not to create or construct a taxonomy of errors. For example, the most detailed classification (although it mainly focuses on parsing aspects) is found in the work of Adriaanse and Rensleigh (2013), whose analysis was based on a sample of only 14 South African environment journals.

In any case, and given their interest, Table VI is a compilation of the main types of errors published to date, the article in which they appeared, and their main items.

4. DISCUSSION AND CONCLUSIONS

The results of our qualitative analysis reveal that the bibliographical corpus on errors in Google Scholar is still limited. The bibliographic review process yielded a total of 49 publications, of which only a

small percentage deals in any depth with the concept of errors and even fewer contribute empirical data.

With the exception of Péter Jacsó's work, we can only point to two articles written with the goal of directly ascertaining how errors in Google Scholar function and what their impact is: Doğan et al. (2016) and Orduna-Malea et al. (2017). Other works of great interest, such as those by Harzing and Van der Wal (2008), Baneyx (2008), Li et al. (2010), Adriaanse and Rensleigh (2011; 2013), and De

Table VI. Error Classifications in Google Scholar

Work	Type	Subtype	Original error Category
Jacsó (2008a)	Innumeracy		Browsing
	Illiteracy		Parsing
Jacsó (2008c)	Phantom link		Browsing
	Phantom master record		Matching
	Phantom author		Parsing
	Phantom citation		Matching
	Phantom publication year		Parsing
Jacsó (2010)	Phantom author	From section title	Parsing
		From search template option	Parsing
		From other data elements	Parsing
	Phantom publication year		Parsing
Adriaanse and Rensleigh (2011)	Data export inconsistencies		Browsing
	Author inconsistencies	Author spelling	Parsing
		Double-barrelled surnames	Parsing
		Diacritical marks	Parsing
		Author sequence	Parsing
		Separated authors	Parsing
		Incorrect authors	Parsing
		Nonsensical authors	Parsing
	Article title inconsistencies	Article title omission	Parsing
		Article title variation	Parsing
		Article title sequence	Parsing
		Subtitle omission	Parsing
		Alternate language title	Parsing
		Nonsensical title	Parsing
	Volume inconsistencies	Volume number omission	Parsing
	Issue inconsistencies	Issue number omissions	Parsing
		Incorrect issue number	Parsing
	Page number inconsistencies	Page number omitted	Parsing
		Incorrect page number	Parsing

Winter et al. (2014), in spite of their contributions to the knowledge of errors in Google Scholar, have addressed this issue in an indirect, circumstantial, secondary or at worst incidental way.

This means that, in general terms, scholarly literature about errors in Google Scholar, particularly articles focusing on the use of this tool in bibliometric analysis, is scarce, excessively fragmented and diffuse. There are no studies in which research designs have been specifically developed not only to identify but also to quantify the errors and evaluate their consequences. Studies that do touch on the question of errors were designed with other objectives in mind, and when they address the issue, they often arrive at conclusions that are all too apparent (that there are errors is obvious). In addition, the few studies that provide empirical evidence (albeit indirectly) are not comparable because they deal with completely different samples with different units and research objectives.

Given the importance of quantifying and evaluating the consequences of errors in Google Scholar, since this database is widely used in both bibliometric analysis and in academic evaluation processes (whether we like it or not), it is quite remarkable that the bibliometric community has not undertaken more studies of this nature. The experts that have been most critical of Google Scholar, with the exception of Jacsó, have criticised the database on the basis of its errors, but have not studied their true impact on bibliometric analysis, especially in the context of a big data system that is forcibly transforming the postulates on which many bibliometric studies have been based. These studies are limited – for better or for worse – by the capabilities of the available bibliographic sources, which to date have been controlled and supervised.

One of the possible reasons is the recognised difficulty in evaluating the errors themselves, due to certain substantial limitations (limit of 1,000 search results, limit of 1,000 citing documents per result, with hardly any options for ordering the results, etc.). This is something that has been strongly criticised by Jacsó (2006a; 2008c; 2012b), while Meho and Yang (2007) have already criticised the excessive time required to clean up the data.

For this reason, few studies have shed light on the real effects of existing errors. Sanderson (2008), who calculated the h-index in detail for 3 British researchers, concluded that, after correcting the errors, the h-index had been underestimated by 5-10%. Li et al. (2010), who also acknowledged the excessive data processing

time required by Google Scholar, showed that data cleaning processes have, after all, little effect on results, something that had already been partially demonstrated by Baneyx (2008), albeit with very small samples. Doğan et al. (2016) were the first to systematically calculate various indicators before and after cleaning the data (in this case in Google Scholar Metrics). Although the authors concluded that the differences in the calculation of the h-index and the i10-index before and after eliminating duplicates (of both records and received citations) were statistically significant, an analysis of their results leads us to question their conclusion, since the differences, even when they exist, are not so significant. In fact, the h-index does not change for any of the authors after deleting the duplicate records, although it does change slightly after deleting duplicate citations (the most extreme case falls from 16 to 13). In these cases, the level of profile editing and maintenance (even possible manipulation) by the authors themselves has a direct influence on these differences.

Lastly, as regards Jacsó's work, his quite considerable body of work identifying, discovering, testing and disseminating the errors and limitations of Google Scholar are worthy of recognition. Undoubtedly, he is the author who has most contributed to the serious, rigorous and non-opinionated analysis of this database, so that it may be used for bibliometric purposes. Nevertheless, we would venture to mention some limitations or shortcomings in his extensive scholarly output. Regrettably, Jacsó's work does not reveal all the errors in Google Scholar, although it does expose the most notorious and flagrant, a fact that has led to an improved service. Many of the errors are perhaps repeated excessively throughout his work as practical examples and, beyond the self-explanatory screenshots, greater detail would not have gone amiss in some of the methodological aspects, which are sometimes lacking or only partly sketched out. The design of an exhaustive systematic classification of errors, as well as an estimation of the overall magnitude of these errors, beyond simple exemplification, is also lacking. This has become particularly relevant since 2012 (when Péter Jacsó's contributions ceased and GSC and GSM appeared on the scene).

The evolution of Google Scholar (both in coverage and data quality) must be continually evaluated because of the speed at which its database is updated. Nevertheless, the tests performed in the course of this study have shown that most of the errors reported by Jacsó (especially parsing and searching errors) are still present today. However, the calculation of bibliometric indicators (citations

received, h-index) has improved, thanks in no small measure to the development and evolution of GSM and GSC (predicted by Jacsó, 2012a). Only the calculation of error rates (by type of error), with large samples and by discipline, will allow us to rigorously appraise the suitability of the system for use as a complement to the evaluation of academic impact.

5. ACKNOWLEDGEMENTS

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

Orduna-Malea holds a postdoctoral fellowship (PAID-10-14), from the Polytechnic University of Valencia (Spain). This manuscript has been translated by professional native translator Charles Balfour.

AGRADECIMIENTOS

Alberto Martín-Martín está en cuarto año de la beca de doctorado (FPU2013/05863) otorgada por el Ministerio de educación Cultura y Deportes (españa). Enrique Orduña-Malea es becario postdoctoral (PAID-10-14) de la Universidad Politécnica de Valencia. (España). Este artículo ha sido traducido por el traductor profesional Charles Balfour.

6. NOTES

1. <https://scholar.google.com/intl/en/scholar/about.html>
2. Like many informal articles published about Google Scholar between 2004 and 2005, this document is no longer available online, and has been retrieved from the Wayback Machine (archive.org). See bibliography.
3. http://www.searchengineshowdown.com/blog/2008/01/scholar_down_books_up.shtml
4. <https://techcrunch.com/2007/12/22/2007-in-numbers-igoogles-googles-homegrown-star-performer-this-year/>
5. Google Scholar Blog, available at: <http://weblogs.elearning.ubc.ca/googlescholar/>
6. <https://scholar.google.com/intl/en/scholar/inclusion.html#content>
7. It should be remembered that the number of citations was the criterion used by Google Scholar to rank the results (Jacsó, 2008b). As of 2007/2008, new ranking criteria (<http://scholar.google.com/intl/en/scholar/about.html>) were introduced, of which citations is only one, though important, criterion (Martín-Martín et al., 2017).
8. <http://wokinfo.com/googlescholar/>
9. <https://www.harzing.com/resources/publish-or-perish>
10. <https://scholar.googleblog.com/2011/11/google-scholar-citations-open-to-all.html>
11. <https://scholar.googleblog.com/2012/04/google-scholar-metrics-for-publications.html>

7. REFERENCES

- Abram, S. (2005). Google Scholar: thin edge of the wedge?. *Information Outlook*, 9 (1), 44-46.
- Adlington, J.; Benda, C. (2006). Checking under the hood: evaluating Google scholar for reference use. *Internet Reference Services Quarterly*, 10 (3/4), 135-148.
- Adriaanse, L.; Rensleigh, C. (2011). Content versus quality: a Web of Science, Scopus and Google Scholar comparison. *13th Annual Conference on World Wide Web applications*, pp. 5-18. Cape Peninsula University of Technology; Johannesburg, South Africa.
- Adriaanse, L.; Rensleigh, C. (2013). Web of Science, Scopus and Google Scholar: A content comprehensiveness comparison. *The Electronic Library*, 31 (6), 727-744. <https://doi.org/10.1108/el-12-2011-0174>
- Aguillo, Isidro F. (2012). Is Google Scholar useful for bibliometrics? A webometric analysis. *Scientometrics*, 91 (2), 343-351. <http://dx.doi.org/10.1007/s11192-011-0582-8>
- Baneyx, A. (2008). "Publish or Perish" as citation metrics used to analyze scientific output in the humanities: International case studies in Economics, Geography, Social Sciences, Philosophy, and History. *Archivum Immunologiae et Therapiae Experimentalis*, 56 (6), 363-371. <https://doi.org/10.1007/s00005-008-0043-0>
- Bar-Ilan, J. (2006). An ego-centric citation analysis of the works of Michael O. Rabin based on multiple citation indexes. *Information Processing & Management*, 42 (6), 1553-1566. <https://doi.org/10.1016/j.ipm.2006.03.019>
- Bar-Ilan, J. (2008). Which h-index?—A comparison of WoS, Scopus and Google Scholar. *Scientometrics*, 74 (2), 257-271. <https://doi.org/10.1007/s11192-008-0216-y>
- Bar-Ilan, J. (2010). Citations to the "Introduction to informetrics" indexed by WOS, Scopus and Google Scholar. *Scientometrics*, 82 (3), 495-506. <https://doi.org/10.1007/s11192-010-0185-9>
- Bauer, K.; Bakkalbasi, N. (2005). An examination of citation counts in a new scholarly communication environment. *D-Lib magazine*, 11 (9). <https://doi.org/10.1045/september2005-bauer>

- Beel, J.; Gipp, B. (2009). Google Scholar's ranking algorithm: an introductory overview. *Proceedings of the 12th international conference on scientometrics and informetrics*, pp. 230-241. ISSI. Rio de Janeiro, Brazil.
- Belew R.K. (2005). Scientific impact quantity and quality: analysis of two sources of bibliographic data. Available at: <http://www.cogsci.ucsd.edu/~rik/papers/belew05-iqq.pdf>
- Bensman, S.J. (2012). The impact factor: its place in Garfield's thought, in science evaluation, and in library collection management. *Scientometrics*, 92 (2), 263-275. <https://doi.org/10.1007/s11192-011-0601-9>
- Bosman, J; Mourik, I; Van Rasch, M; Sieverts, E; Verhoeff, H (2006). *Scopus reviewed and compared. The coverage and functionality of the citation database Scopus, including comparisons with Web of Science and Google Scholar*. Utrecht University Library. Available at: <https://dspace.library.uu.nl/handle/1874/18247>
- Breeding, M. (2015). *The future of library resource discovery*. NISO Whitepapers. NISO; Baltimore, United States.
- Butler, D. (2004). Science searches shift up a gear as Google starts Scholar Engine. *Nature*, 432, p. 423. <https://doi.org/10.1038/432423a>
- Butler, L. (2011). The devil is in the detail: Concerns about Vanclay's analysis of Australian journal rankings. *Journal of Informetrics*, 5 (4), 693-694. <https://doi.org/10.1016/j.joi.2011.04.001>
- De Winter, J.C.; Zadpoor, A.A.; Dodou, D. (2014). The expansion of Google Scholar versus Web of Science: a longitudinal study. *Scientometrics*, 98 (2), 1547-1565. <https://doi.org/10.1007/s11192-013-1089-2>
- Dilger, A.; Müller, H. (2013). A citation-based ranking of German-speaking researchers in business administration with data of Google Scholar. *European Journal of Higher Education*, 3 (2), 140-150. <https://doi.org/10.1080/21568235.2013.779464>
- Doğan, G.; Şencan, İ.; Tonta, Y. (2016). Does dirty data affect google scholar citations?. *Proceedings of the Association for Information Science and Technology*, 53 (1), 1-4. <https://doi.org/10.1002/pra2.2016.14505301098>
- Felter, L.M. (2005). The better mousetrap: Google Scholar, Scirus, and the Scholarly Search Revolution, *Searcher*, 13 (2), 43-48.
- García-Pérez, M.A. (2010). Accuracy and completeness of publication and citation records in the Web of Science, PsycINFO, and Google Scholar: A case study for the computation of h indices in Psychology. *Journal of the Association for Information Science and Technology*, 61(10), 2070-2085. <https://doi.org/10.1002/asi.21372>
- Gardner, S.; Eng, S. (2005). Gaga over Google? Scholar in the social sciences. *Library Hi Tech News*, 22 (8), 42-45. <https://doi.org/10.1108/07419050510633952>
- Giles, J. (2005). Science in the web age: Start your engines. *Nature*, 438 (7068), 554-555. <https://doi.org/10.1038/438554a>
- Goodman, A. (2004). Google Scholar vs. Real Scholarship, *Traffic*. Available at: <http://www.traffick.com/2004/11/google-scholar-vs-real-scholarship.asp>
- Haddaway, N.R.; Collins, A.M.; Coughlin, D.; Kirk, S. (2015). The role of Google Scholar in evidence reviews and its applicability to grey literature searching. *PLoS one*, 10 (9), e0138237. <https://doi.org/10.1371/journal.pone.0138237>
- Harzing, A.W. (2010). *The publish or perish book*. Tarma software research; Melbourne.
- Harzing, A.W. (2014). A longitudinal study of Google Scholar coverage between 2012 and 2013. *Scientometrics*, 98 (1), 565-575. <https://doi.org/10.1007/s11192-013-0975-y>
- Harzing, A.W.; Alakangas, S. (2016). Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison. *Scientometrics*, 106 (2), 787-804. <https://doi.org/10.1007/s11192-015-1798-9>
- Harzing, A.W.; Van der Wal, R. (2008). Google Scholar as a new source for citation analysis. *Ethics in science and environmental politics*, 8 (1), 61-73. <https://doi.org/10.3354/esepp00076>
- Hirsch, J.E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102 (46), 16569-16572. <https://doi.org/10.1073/pnas.0507655102>
- Jacsó, P. (2004). Péter's digital ready reference shelf, (web-only document). Available at: <https://goo.gl/ouV3PP>
- Jacsó, P. (2005a). As we may search: Comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases. *Current science*, 89 (9), 1537-1547.
- Jacsó, P. (2005b). Comparison and analysis of the citedness scores in Web of Science and Google Scholar. *International Conference on Asian Digital Libraries*, pp 360-369. Springer; Berlin; Heidelberg, Germany. https://doi.org/10.1007/11599517_41
- Jacsó, P. (2005c). Google Scholar: the pros and the cons. *Online information review*, 29 (2), 208-214. <http://dx.doi.org/10.1108/14684520510598066>
- Jacsó, P. (2006a). Deflated, inflated, and phantom citation counts. *Online Information Review*, 30 (3), 297-309. <http://dx.doi.org/10.1108/14684520610675816>
- Jacsó, P. (2006b). Dubious hit counts and cuckoo's eggs. *Online Information Review*, 30 (2), 188-193. <https://doi.org/10.1108/14684520610659201>
- Jacsó, P. (2008a). Google scholar revisited. *Online Information Review*, 32 (1), 102-114. <https://doi.org/10.1108/14684520810866010>
- Jacsó, P. (2008b). The pros and cons of computing the h-index using Google Scholar. *Online Information Review*, 32 (3), 437-452. <http://dx.doi.org/10.1108/14684520810889718>

- Jacsó, P. (2008c). Testing the calculation of a realistic h-index in Google Scholar, Scopus, and Web of Science for F.W. Lancaster. *Library Trends*, 56 (4), 784-815. <https://doi.org/10.1353/lib.0.0011>
- Jacsó, P. (2009a). Calculating the h-index and other bibliometric and scientometric indicators from Google Scholar with the Publish or Perish software. *Online Information Review*, 33(6), 1189-1200. <https://doi.org/10.1108/14684520911011070>
- Jacsó, P. (2009b). Google Scholar's Ghost Authors. *Library Journal*, 134 (18), 26-27.
- Jacsó, P. (2010). Metadata mega mess in Google Scholar. *Online Information Review*, 34 (1), 175-191. <https://doi.org/10.1108/14684521011024191>
- Jacsó, P. (2011). Google Scholar duped and deduped—the aura of “robometrics”. *Online Information Review*, 35(1), 154-160. <https://doi.org/10.1108/14684521111113632>
- Jacsó, P. (2012a). Google Scholar Author Citation Tracker: is it too little, too late?. *Online Information Review*, 36(1), 126-141. <https://doi.org/10.1108/14684521211209581>
- Jacsó, P. (2012b). Grim tales about the impact factor and the h-index in the Web of Science and the Journal Citation Reports databases: Reflections on Vanclay's criticism. *Scientometrics*, 92 (2), 325-354. <https://doi.org/10.1007/s11192-012-0769-7>
- Jacsó, P. (2012c). Using Google Scholar for journal impact factors and the h-index in nationwide publishing assessments in academia – siren songs and air-raid sirens. *Online Information Review*, 36 (3), 462-478. <http://dx.doi.org/10.1108/14684521211241503>
- Jacsó, P. (2012d). Google Scholar Metrics for Publications: The software and content features of a new open access bibliometric service. *Online Information Review*, 36 (4), 604-619. <https://doi.org/10.1108/14684521211254121>
- Kennedy, S.; Price, G. (2004). Big News: “Google Scholar” is Born. *ResourceShelf*. Available at: <http://web.resourceshelf.com/go/resourceblog/40511>
- Leslie M.A. (2004). A Google for academia. *Science*, 306 (5702), 1661-1663. <https://doi.org/10.1126/science.306.5702.1661c>
- Levine-Clark, M.; Gil, E.L. (2009). A comparative citation analysis of Web of Science, Scopus and Google Scholar. *Journal of Business and Finance Librarianship*, 14 (1), 32-46. <https://doi.org/10.1080/08963560802176348>
- Li, J.; Sanderson, M.; Willett, P.; Norris, M.; Oppenheim, C. (2010). Ranking of library and information science researchers: Comparison of data sources for correlating citation data, and expert judgments. *Journal of Informetrics*, 4 (4), 554-563. <https://doi.org/10.1016/j.joi.2010.06.005>
- London School of Economics and Political Science (2011). *Maximizing the impacts of your research: A handbook for social scientists*. LSE; UK. Available at: http://www2.lse.ac.uk/government/research/resgroups/LSEPublicPolicy/Docs/LSE_Impact_Handbook_April_2011.pdf
- Maia, J.L.; Di Serio, L.C.; Alves Filho, A.G. (2016). Bibliometric research on strategy as practice: exploratory results and source comparison. *Sistemas & Gestão*, 10 (4), 654-669. <https://doi.org/10.20985/1980-5160.2015.v10n4.662>
- Martín-Martín, A.; Ayllón, J.M.; Orduna-Malea, E.; Delgado López-Cózar, E. (2014a). Google Scholar Metrics 2014: a low cost bibliometric tool. *EC3 Working Papers*, 17. Available at: <https://arxiv.org/abs/1407.2827>
- Martín-Martín, A., Orduna-Malea, E., Ayllón, J.M.; Delgado López-Cózar, E. (2014b). Does Google Scholar contain all highly cited documents (1950-2013)?. *EC3 Working Papers*, 19. Available at: <https://arxiv.org/abs/1410.8464>
- Martín-Martín, A.; Ayllón, J.M.; Orduna-Malea, E.; Delgado López-Cózar, E. (2016a). 2016 Google Scholar Metrics released: a matter of languages... and something else. *EC3 Working Papers*, 22. Available at: <https://arxiv.org/abs/1607.06260>
- Martín-Martín, A.; Orduna-Malea, E.; Ayllón, J.M.; Delgado López-Cózar, E. (2016b). A two-sided academic landscape: snapshot of highly-cited documents in Google Scholar (1950-2013). *Revista Española de Documentación Científica*, 39 (4). <http://dx.doi.org/10.3989/redc.2016.4.1405>
- Martín-Martín, A.; Orduna-Malea, E.; Ayllón, J.M.; Delgado López-Cózar, E. (2016c). The counting house: measuring those who count. Presence of Bibliometrics, Scientometrics, Informetrics, Webometrics and Altmetrics in the Google Scholar Citations, ResearcherID, ResearchGate, Mendeley & Twitter. *EC3 Working Papers*, 21. Available at: <https://arxiv.org/abs/1602.02412>
- Martín-Martín, A.; Orduna-Malea, E.; Harzing, A.W.; Delgado López-Cózar, E. (2017). Can we use Google Scholar to identify highly-cited documents?. *Journal of Informetrics*, 11 (1), 152-163. <https://doi.org/10.1016/j.joi.2016.11.008>
- Meho, L.I.; Yang, K. (2007). Impact of data sources on citation counts and rankings of LIS faculty: Web of Science versus Scopus and Google Scholar. *Journal of the American Society for Information Science and Technology*, 58 (13), 2105-2125. <https://doi.org/10.1002/asi.20677>
- Moed, H.F.; Bar-Ilan, J.; Halevi, G. (2016). A new methodology for comparing Google Scholar and Scopus. *Journal of Informetrics*, 10 (2), 533-551. <https://doi.org/10.1016/j.joi.2016.04.017>
- Noll, H.M. (2008). *Where Google Scholar Stands on Art: An Evaluation of Content Coverage in Online Databases*. [Master Thesis]. University of North Carolina at Chapel Hill; North Carolina.
- Noruzi, A. (2005). Google Scholar: the new generation of citation indexes. *Libri*, 55 (4), 170-180. <https://doi.org/10.1515/libr.2005.170>

- Notess, G.R. (2005). Scholarly web searching: Google Scholar and Scirus. *Online*, 29 (4), 39-41.
- Nunberg, G. (2009). Google's book search: A disaster for scholars. *The chronicle of higher education*, 31. Available at: <http://www.chronicle.com/article/Googles-Book-Search-A/48245>
- Oder, N. (2009). Google, 'the last library', and millions of metadata mistakes. *Library Journal Academic Newswire*, 3.
- Ojala, M. (2005). Scholarly mistakes. *Online*, 29 (3), 26.
- Orduna-Malea, E.; Martín-Martín, A.; Ayllón, J.M.; Delgado López-Cózar, E. (2016). *La revolución Google Scholar. Destapando la caja de Pandora académica*. UNE (Unión de Editoriales Universitarias Españolas); Granada.
- Orduna-Malea, E.; Ayllón, J.M.; Martín-Martín, A.; Delgado López-Cózar, E. (2017). The lost academic home: institutional affiliation links in Google Scholar Citations. *Online Information Review*, 41 (6), 762-781. <https://doi.org/10.1108/OIR-10-2016-0302>
- Ortega, J. L. (2014). *Academic search engines: A quantitative outlook*. Elsevier; Oxford. <http://www.sciencedirect.com/science/book/9781843347910>
- Ortega, J. L. (2015). Relationship between altmetric and bibliometric indicators across academic social sites: The case of CSIC's members. *Journal of Informetrics*, 9 (1), 39-49. <https://doi.org/10.1016/j.joi.2014.11.004>
- Pauly, D.; Stergiou, K.I. (2005). Equivalence of results from two citation analyses: Thomson ISI's Citation Index and Google's Scholar service. *Ethics in Science and Environmental Politics*, 9, 33-35. <https://doi.org/10.3354/esep005033>
- Perkel (2005). The future of citation analysis. *The Scientist*, 19 (20), 24.
- Pitol, S.P.; De Groote, S.L. (2014). Google Scholar versions: do more versions of an article mean greater impact?. *Library Hi Tech*, 32 (4), 594-611. <https://doi.org/10.1108/lht-05-2014-0039>
- Price, G. (2004). Google Scholar documentation and large PDF files. *Search Engine Watch*. Available at: <https://searchenginewatch.com/sew/news/2063361/google-scholar-documentation-large-pdf-files>
- Robinson, M.L.; Wusteman, J. (2007). Putting Google Scholar to the test: A preliminary study. *Program*, 41 (1), 71-80. <https://doi.org/10.1108/00330330710724908>
- Rosenstreich, D.; Wooliscroft, B. (2009). Measuring the impact of accounting journals using Google Scholar and the g-index. *The British Accounting Review*, 41 (4), 227-239. <https://doi.org/10.1016/j.bar.2009.10.002>
- Sanderson, M. (2008). Revisiting h measured on UK LIS and IR academics. *Journal of the American Society for Information Science and Technology*, 59 (7), 1184-1190. <https://doi.org/10.1002/asi.20771>
- Shultz M. (2007). Comparing test searches in PubMed and Google Scholar. *Journal of the Medical Library Association*, 95 (4), 442-445. <https://doi.org/10.3163/1536-5050.95.4.442>
- Sullivan, D. (2004). Google Scholar Offers Access to Academic Information. *Search Engine Watch*. Available at: <https://searchenginewatch.com/sew/news/2048646/google-scholar-offers-access-to-academic-information>
- Thelwall, M.; Kousha, K. (2017). ResearchGate versus Google Scholar: Which finds more early citations?. *Scientometrics*, 112 (2), 1125-1131. <https://doi.org/10.1007/s11192-017-2400-4>
- Thor, A.; Bornmann, L. (2011). The calculation of the single publication h index and related performance measures: A web application based on Google Scholar data. *Online Information Review*, 35 (2), 291-300. <https://doi.org/10.1108/14684521111128050>
- Torres-Salinas, D.; Ruiz-Pérez, R.; Delgado-López-Cózar, E. (2009). Google Scholar como herramienta para la evaluación científica. *El profesional de la información*, 18 (5), 501-510. <https://doi.org/10.3145/epi.2009.sep.03>
- Vanclay, J.K. (2012). Impact factor: outdated artefact or stepping-stone to journal certification?. *Scientometrics*, 92 (2), 211-238. <https://doi.org/10.1007/s11192-011-0561-0>
- Vaughan, L.; Shaw, D. (2008). A New Look at Evidence of Scholarly Citations in Citation Indexes and From Web Sources. *Scientometrics*, 74 (2), 317-330. <https://doi.org/10.1007/s11192-008-0220-2>
- Verstak, A.; Acharya, A. (2013). Identifying multiple versions of documents. *US Patents* (US8589784 B1). Available at: <https://www.google.com/patents/US8589784>
- Vine, R. (2005). Google Scholar is a full year late indexing Pubmed content. *SiteLines: ideas about searching*. Available at: http://web.archive.org/web/20060716085124/http://www.workingfaster.com/sitelines/archives/2005_02.html
- Walters, W.H. (2007). Google Scholar coverage of a multidisciplinary field. *Information Processing & Management*, 43 (4), 1121-1132. <https://doi.org/10.1016/j.ipm.2006.08.006>
- White, B. (2006). Examining the claims of Google Scholar as a serious information source. *New Zealand Library & Information Management Journal*, 50 (1), 11-24.
- Wlekinski, J.M. (2005). Studying Google Scholar: wall to wall coverage?. *Online*, 29 (3), 22-26.
- Yang, K.; Meho, L.I. (2006). Citation analysis: a comparison of Google Scholar, Scopus, and Web of Science. *Proceedings of the American Society for information science and technology*, 43 (1), 1-15. <https://doi.org/10.1002/meet.14504301185>

APPENDIX I. PÉTER JACSÓ'S WORK ON GOOGLE SCHOLAR

WORK	ANALYSIS UNIT	SAMPLE	Period	DISCIPLINE	GEOGRAPHIC COVERAGE	ANALYSED PRODUCT	Main variable	Empirical Observations
Jacsó (2004)	Hit results	Random Test searches	All	Multidisciplinary	International	Google Scholar	-Searching -Linking -Citations -Versions	-Deficiencies are found using "site" command -Hits and citations inflated -Versions unmerged
Jacsó (2005a)	Hit results Authors (Garfield) Journals (Current science) Documents	Random Test searches	All to May 2005	Multidisciplinary	International	Google Scholar	-Searching -Parsing	-Deficiencies in year range searches -Deficiencies in Boolean searches -Deficiencies in advanced search (published by function) -Deficiencies in citation attribution
Jacsó (2005b)	Journal	680 articles published in Asian Pacific Journal of Allergy and Immunology	December 2004 – July 2005	Medicine	International	Google Scholar	-Parsing -Matching	-Deficiencies in the correct indexation of one journal related especially to author-ity control (source title variants, article title variants), erroneous titles, and article versions unmerged
Jacsó (2005c)	Journal	Random Test searches		Multidisciplinary	International	Google Scholar	-Searching	-Deficiencies in coverage using "site" search command
Jacsó (2006a)	Hit results	Random Test searches	2005; 2006	Multidisciplinary	International	Google Scholar	-Searching -Matching	-Through the use of one specific query search (<scholars>, the number of hit counts retrieved are analysed as well as the number of citations received by each result. Some errors are found (false positives) which inflate results. -Cited/citing document linkage malfunction generates citation errors -Errors in the bibliographic information (authorship, date of publication, source) are found
Jacsó (2006b)	Hit results	Random Test searches		Multidisciplinary	International	Google Scholar	-Searching	-Duplicate hits -Less than 1,000 hits against some queries. -Boolean logic does not work properly in combined query searches

Jacsó (2008a)	Hit results	Random Test searches		Multidisciplinary	International	Google Scholar	-Searching -Parsing	-Limitations and errors in GS are classified into Innumeracy and Illiteracy -Errors are compiled and recalculated to compare them to those previously detected in 2007
Jacsó (2008b)	Authors Journals	Random Test searches		Multidisciplinary	International	Google Scholar	-Searching -Parsing	-Google Scholar (GS) is under the microscope from the perspective of calculating the h-index for individuals and journals. -Phantom citations -Duplicate hits -Identification of Ghost authors that steal citations from real authors -Hit count estimates vary according to each SERP
Jacsó (2008c)	Authors	Random Test searches		Multidisciplinary	International	Google Scholar	-Searching -Parsing	-Discusses the related features of Google Scholar, Scopus, and Web of Science (WoS), and demonstrates in the latter how a much more realistic and fair h-index can be computed for F. W. Lancaster than one produced automatically -Categorisation of errors: phantom links, phantom master records, phantom authors, phantom citations, phantom publication years
Jacsó (2009a)	Authors Journals	Random Test searches		Multidisciplinary	International	Google Scholar	-Searching -Matching	-Calculating errors in GS when using Pop software, especially when computing h-index -Summarising errors with Boolean searches, duplicates (hits and citations), and author mismatches
Jacsó (2010)	Authors	Random Test searches		Multidisciplinary	International	Google Scholar	-Parsing	-Phantom authors -Phantom publication years -Critique of GS team feedback
Jacsó (2011)	Authors	Random Test searches		Multidisciplinary	International	Google Scholar	-Coverage	-Some of the problems that exist with GS are discussed, particularly regarding content spam and citation spam

Jacsó (2012a)	Authors	1 author (A. Acharya), developer of GS		Multidisciplinary	International	Google Scholar Citations	-Authors	-Improvements in GS data are highlighted: elimination of duplicates, and availability of bibliometric indicators. -Some errors are equally claimed: co-author lists, citations (inherited from GS) and search limitations
Jacsó (2012b)	Journals	Random Test searches		Forestry	International	Google Scholar	-Searching -Parsing	-Critique of Vanclay's work on using Google Scholar's h-index instead of WoS and JCR; methodological errors and misunderstood concepts are highlighted as well
Jacsó (2012c)	Hit results Journals	Random Test searches	-	Forestry	International	Google Scholar	-Searching -Parsing	-Critique of Vanclay's work. In this case, the Google Scholar ranking algorithm is observed, as well as some other related limitations (1,000 results per query, lack of advanced sorting results options) which jeopardise the use of GS to accurately compute some basic bibliometric indicators, irrespective of the existence of other errors.
Jacsó (2012d)	Journals	20 journals	-	Marketing	International	Google Scholar Metrics	-Searching -Parsing	-The purpose of this paper is to review the software and content features of the Google Scholar Metrics (GSM) service launched in April 2012, examining the software, browsing, searching and sorting functions, citation matching and content.

APPENDIX II. COMPILATION OF STUDIES THAT DIRECTLY OR INDIRECTLY ADDRESS ERRORS IN GOOGLE SCHOLAR

WORK	ANALYSIS UNIT	SAMPLE	Period	DISCIPLINE	GEOGRAPHIC COVERAGE	PRODUCT	Variables	Empirical Observations
Belew (2005)	Authors	203 publications from 6 authors	1997-2004 (publication date of cited works)	Multidisciplinary	International	GS	<ul style="list-style-type: none"> • Improper citation attribution • Author inconsistencies 	<ul style="list-style-type: none"> • 36 citations identified by WoS but missed by GS
Bar-Ilan (2006)	Documents	113 publications of Michael Rabin	N/A	Mathematics; Computer Science	International	GS	<ul style="list-style-type: none"> • Improper citation attribution • Improper author attribution • Author inconsistencies 	<ul style="list-style-type: none"> • No empirical data on errors is provided
Yang and Meho (2006)	Authors	47 and 62 publications from 2 authors	N/A	Library & Information Science	USA	GS	<ul style="list-style-type: none"> • Duplicate citations 	<ul style="list-style-type: none"> • No empirical data on errors is provided
Meho and Yang (2007)	Authors	5,285 citations from 25 researchers	1971-2005	Library & Information Science	USA	GS	<ul style="list-style-type: none"> • Author inconsistencies • Duplicate citations • Matching errors 	<ul style="list-style-type: none"> • GS misses 501 (12%) citations due to system errors • 475 citations from GS did not have complete bibliographic information
Robinson and Wusteman (2007)	Hit results	7 students through 8 queries	N/A	Multidisciplinary	Ireland	GS	<ul style="list-style-type: none"> • Duplicate hits 	<ul style="list-style-type: none"> • 2 queries with 1 inaccessible result each
Shultz (2007)	Hit results	10 queries	2006	Medicine	International	GS	<ul style="list-style-type: none"> • Duplicate hits • Pages not found 	<ul style="list-style-type: none"> • Duplicates in 7 out of 10 queries • 4 queries out of 10 with some error • 5 pages with errors out of 247 results
Walters (2007)	Documents	155 articles	1990-2000	Later-life migration	International	GS	<ul style="list-style-type: none"> • Duplicate publications • Author inconsistencies • Title inconsistencies 	<ul style="list-style-type: none"> • 1 citation error (a misspelled name) and 1 link that led to the wrong destination page
Baneyx (2008)	Authors	12 researchers	N/A	History; Sociology; Economy; Geography; Philosophy	France	GS	<ul style="list-style-type: none"> • Duplicate publications 	<ul style="list-style-type: none"> • 86% of citations found by Pop are incorrect
Bar-Ilan (2008)	Authors	40 highly-cited Israeli researchers	1996-2006	Multidisciplinary	Israel	GS	<ul style="list-style-type: none"> • Error in date of publication • Search errors • Versions not properly grouped • Wrong authorship • Similar titles are merged wrongly 	<ul style="list-style-type: none"> • "Probabilistic encryption" cited 915 times improperly attributed to A. Wigderson • Duplications to "P Ingwersen" query • Date of publication for American Physical Society fails frequently

Harzing and Van de Wal (2008)	Journals	20	2002-2006	Management and International Business	International	GS GSM	<ul style="list-style-type: none"> •Author inconsistencies •Title inconsistencies •Citations improperly attributed 	<ul style="list-style-type: none"> •Query for "I Introduction" shows 956 results, of which 160 correspond to publications a) that receive citations, and b) of which "I Introduction" is the only author
Noll (2008)	Documents	472	1975-2008	History of Art	International	GS	<ul style="list-style-type: none"> •Duplicate citations •Duplicate publications •Incomplete bibliographic information 	<ul style="list-style-type: none"> •23% of citations included duplicates and multiple versions
Rosenstreich and Wooliscro (2009)	Journals	34	N/A	Accounting	International	GS	<ul style="list-style-type: none"> •Title inconsistencies •Duplicate publications •Duplicate citations 	<ul style="list-style-type: none"> •Overall, the level of duplicates within GS was found to be around 3%
Bar-Ilan (2010)	Documents	1 book: 'introduction to informetrics' (by egghe and rousseau)	N/A	Informetrics; Bibliometrics	International	GS	<ul style="list-style-type: none"> •Title inconsistencies •Duplicate publications •Duplicate citations •Versions unmerged 	<ul style="list-style-type: none"> •24 items out of 358 (6.7%) referring to the book are duplicates (17 title duplicates and 5 with authorship errors) •First SERP: Hit count estimates claims 227 results, but only 219 appear •Second SERP: 134 results are claimed, and only 119 appear. •After deleting duplicates, 51 citations (14.2%) constitute errors.
García-Pérez (2010)	Authors	380 publications from 4 authors	N/A	Psychology	Spain	GS	<ul style="list-style-type: none"> •Duplicate citations •Duplicate publications •Citations missed •Phantom citations •Title inconsistencies 	<ul style="list-style-type: none"> •55 citations (16.5%) present errors •11 errors were due to HCE •57 errors due to phantom citations •215 contains duplicate links •100 duplicates due to reprints published in different sources
Li et al. (2010)	Authors	42 experts rating research published by 101 authors	N/A	Library & Information Science	International	GS	<ul style="list-style-type: none"> •Duplicate publications •Duplicate citations 	<ul style="list-style-type: none"> •Citation counts inflated by 5% after cleaning data through PoP
Adriaanse and Rensleigh (2011)	Journals	9	2004-2008	Environmental Sciences	South Africa	GS	<ul style="list-style-type: none"> •Author inconsistencies •Title inconsistencies •Bibliographic inconsistencies (volume, issue, pagination) 	<ul style="list-style-type: none"> •185 duplicate records •448 (14%) inconsistencies

LSE Public Policy Group (2011)	Authors	120	N/A	Economy; Geography; International relations; Law; Sociology	UK	GS	<ul style="list-style-type: none"> •Duplicate citations •Duplicate publications 	<ul style="list-style-type: none"> •No empirical data on errors is provided
Thor and Bornmann (2011)	Journals	N/A	N/A	Multidisciplinary	International	GS	<ul style="list-style-type: none"> •Duplicate publications 	<ul style="list-style-type: none"> •The query "allintitle: merge purge large" shows 8 results (records) all of which refer to the same publication (duplicates)
Adriaanse and Rensleigh (2013)	Journals	14	2004-2008	Environmental Sciences	South Africa	GS	<ul style="list-style-type: none"> •Author inconsistencies •Title inconsistencies •Bibliographic inconsistencies (volume, issue, pagination) 	<ul style="list-style-type: none"> •448 (14%) inconsistencies: 12 data export inconsistencies; 146 author inconsistencies; 122 title inconsistencies; 167 bibliographic inconsistencies
Dilger and Müller (2013)	Authors	1572 members of German Academic Association for Business Research (VHB)	2005-2009	Business	Germany	GS	<ul style="list-style-type: none"> •Author inconsistencies •Data export inconsistencies •Versions unmerged 	<ul style="list-style-type: none"> •No empirical data on errors is provided
De Winter et al. (2014)	Documents	2 classic articles 56 articles	Sep 2005- Apr 2013	Multidisciplinary	International	GS	<ul style="list-style-type: none"> •Duplicate publications •Duplicate citations •False positives •Technical scanning errors 	<ul style="list-style-type: none"> •8.7% (108) of citations without year of publication. •4% (29) of publications without year of publication (11 in theses; 14 in unknown documents) •64 duplicates and 11 false positives
Harzing (2014)	Authors	20 Nobel winners	2012-2013	Multidisciplinary (Chemistry, Economics; Medicine and Physics)	International	GS	<ul style="list-style-type: none"> •Technical errors 	<ul style="list-style-type: none"> •Ada Yonath lost 1 highly-cited article, modifying her g-index due to a parsing error
Martín-Martín et al. (2014b)	Documents	64000	1950-2013	Multidisciplinary	International	GS	<ul style="list-style-type: none"> •Duplicate publications •Duplicate citations •Title inconsistencies •Phantom authors •Phantom citations •Book reviews confused •Documents improperly attributed to authors •Versions unmerged •Citations improperly attributed to documents •Full text links accessing to erroneous file •Link connection between GS and WoS broken 	<ul style="list-style-type: none"> •29% of documents with undefined typology •165 versions corresponding to one same work unmerged •HCE claims 763 versions for 1 record, but 20.1% are incorrect: 83 broken links, and 38 errors could not be verified

Pitol and De Groot (2014)	Articles	982	N/A	Multidisciplinary	International	GS	<ul style="list-style-type: none"> • Duplicate publications • Versions unmerged 	<ul style="list-style-type: none"> • Not clear if duplicate versions refer to unmerged versions or just merged versions
Haddaway et al. (2015)	Documents	6988 full text searches 4194 Title searches; 7 systematic reviews	N/A	Environmental Sciences	International	GS	<ul style="list-style-type: none"> • Title inconsistencies • Duplicate hits in searches • Data export inconsistencies 	<ul style="list-style-type: none"> • Full text search: duplication rate of 0.56% • Title search: duplication rate of 2.93% • 15% of typographic errors • Capital letters provoke 36% of all duplicates due to Title errors • Incomplete titles cause 15% of duplicates due to title errors • Parsing causes 3% of duplicates due to Title error • Scanning references causes 1.3% of duplicates due to Title error • Total duplication rate: 5%
Ortega (2015)	Authors	191858	2011-2012	Multidisciplinary	International	GSC	<ul style="list-style-type: none"> • Duplicate profiles 	<ul style="list-style-type: none"> • 1.3% (17) duplicate profiles • [duplicate profiles is not considered error but a limitation in this work]
Harzing and Alakangas (2016)	Authors	146	2013-2015	Multidisciplinary (Humanities; Social Sciences; Engineering; Life Sciences)	International	GS	<ul style="list-style-type: none"> • Author inconsistencies • Citation inconsistencies • Duplicate publications 	<ul style="list-style-type: none"> • The use of PoP reduces research output from 244 to 106 results (138 original document duplicates)
Maia et al. (2016)	Publications	2400	N/A	Strategy as Practice (SAP)	International	GS	<ul style="list-style-type: none"> • Author inconsistencies • Incorrect source identification 	<ul style="list-style-type: none"> • 15% (360) documents without date of publication • 27% (633) documents without source of publication

Martín-Martín et al. (2016c)	Authors	814	N/A	Bibliometrics	International	GS GSC	<ul style="list-style-type: none"> •GS: <ul style="list-style-type: none"> •Duplicate publications •Duplicate citations •Title inconsistencies •Phantom authors •Phantom citations •Citation matching •Book reviews misattributed as books •Documents improperly attributed to authors •Versions unmerged •Citations improperly attributed to documents •GSM: <ul style="list-style-type: none"> •Duplicate profiles (deprecated) •Misattributed documents •Data manipulation •Duplicate records •Incorrectly merged documents •Incorrect titles 	<ul style="list-style-type: none"> •No empirical data on errors is provided
Moed et al. (2016)	Journals	1200 articles published in 12 journals	2010-2014	Multidisciplinary	International	GS GSM	<ul style="list-style-type: none"> •Duplicate publications •Duplicate citations •Versions unmerged •Data export inconsistencies •Technical errors 	<ul style="list-style-type: none"> •Article published in Journal of Virology receives 270 citations from GS, out of which 180 are false due to wrong attribution •Total duplicates: 3.9% (430) in GS; 5.2% (366) in GSM. Identical duplicates: 0.2% (26) in GS, and 18% (0.3) in GSM •11.4% of citing documents in GS the publication year is unavailable
Doğan et al. (2016)	Authors	11 authors 617 publications from GS 3144 publications from GSC	N/A	Information management	Turkey	GS GSC	<ul style="list-style-type: none"> •Duplicate publications •Duplicate citations 	<ul style="list-style-type: none"> •73% (8) of authors have duplicate records in GSM profile •55% (135) of publications receive 364 duplicate citations (12% of all citations) •In GS, 92 out of 591 publications are duplicated •In GS, 364 out of 3079 (12%) citations are duplicates

Martín-Martín et al. (2017)	Documents	64000	1950-2013	International	GS	<ul style="list-style-type: none"> •Incorrect title •Versions unmerged •Date of publication inconsistencies 	<ul style="list-style-type: none"> •96.7% of documents included both in WoS and GS (32680) show concordance in date of publication
Orduna-Malea et al. (2017)	Universities Authors	82 universities Sample A: 17938 authors Sample B: 22285 authors	N/A	Spain	GSM	<ul style="list-style-type: none"> •Institution name disambiguation •Incorrect linking •Lack of precision in search queries •Errors in complex institutions due to internal units 	<ul style="list-style-type: none"> •16% (13) Spanish universities do not have a normalised entry •GSC only groups 73.7% of all the scholars that could potentially have been linked to the institution by having an email or affiliation of the institution, and 80.5% of those with official email from the institution