THE UNCONTROLLED GIANT: GOOGLE SCHOLAR & GOOGLE SCHOLAR CITATIONS

The errors that can compromise the metric portrait of an author offered by Google Scholar can be grouped into two main sections. First, the errors Google Scholar sometimes makes when it indexes a document or when it assigns citations to it. Second, the specific errors that are sometimes made during the creation of a Google Scholar Citations profile.

The former are a logical consequence of the tricky and complex task that is automatically searching the current academic papers available in the net. This task also involves merging in only one record all possible versions of the same work, and linking to it all documents in which it is cited (keeping in mind that these documents and references can be presented in the most varied formats).

The latter are the ultimate responsibility of the author, who must periodically revise his/her profile in order to eliminate misattributed documents which might been included in the automatic weekly updates, clean the records by merging different versions of the same document when Google Scholar's algorithms are not able to detect their similarity, as well as improve and complete the bibliographic references of these documents (filling in blank fields in a document when Google Scholar hasn't been able to find that information).

1. **Mistakes in Google Scholar by automatic indexes a document or assigns citations to it.**

Next, we classify, describe, and illustrate some of the most common mistakes in Google Scholar:

a) Incorrect identification of the title of the document

Google Scholar always tries to extract bibliographic information from the HTML Meta tags in a webpage. When there are no Meta tags available, it parses the webpage itself (the HTML code of the page, or even PDFs themselves). Even though its spiders are able to successfully parse pages with a quite broad range of different structures, and despite the fact that they have published a very clear set of inclusion guidelines, some parsing errors occasionally arise for documents extracted from websites with unusual layouts. It is not rare in these cases that an incorrect text string is selected as the title of the document. In Figure 1 we illustrate an example in which an incorrect string ("www.redalyc.org") has been selected as the title of the document in several records, probably because it is the string that is featured with a higher font size in the first page of the PDF document from which Google Scholar has parsed the bibliographic information. Note that the authors and the source publications are correctly assigned.

Figure 1. Document titles improperly identified in Google Scholar: URLs

In many other occasions, other text strings, such as the author's name and/or the year of publication, are incorrectly selected as the title of the document. In Figure 2 we can observe how "de Solla" has been selected as the title in many records.

Figure 2. Author names incorrectly selected as document titles in Google Scholar

https://scholar.google.com/scholar?start=0&q=allintitle:+%22de+solla%22+-Moravcsik+-gulls+-comments+-1922+-foreword+-Toward+-tribute+-space+-pensamento+-address+-appreciation&hl=en&as_sdt=0,5

b) Ghost authors

The topic of ghost authors, citations, and documents was approached by Jacsó in numerous works, mostly before Google Scholar Citations was launched. Although profiles have served to filter and correct many mistakes, some of them still persist, especially if authors do not clean their personal profiles. In Figure 3 we can see one such example. In this case, the record only displays one person as the author of the article (Carmen Martín Moreno), when in fact the article was written by two authors (Elías Sanz-Casado and Carmen Martín Moreno). In this case, Google Scholar extracted the bibliographic information from the HTML Meta tags in the website of the journal where the article was published, but, as we can see in Figure 3 (bottom image), these metadata were already incorrect (the title should read "Técnicas bibliométricas aplicadas a los estudios de usuarios"), and incomplete (Elías Sanz-Casado is missing from the record). Nonetheless, thanks to Google Scholar Citations, Elías was able to add the document to his profile, even if his name is still missing from the authors field (Figure 3, top left).

Figure 3. Missing authors in primary versions of documents in Google Scholar

c) Book reviews indexed as books

Among the most common mistakes in document identification is mistaking the review of a book for the book itself. In Figure 4 we show two different records which correspond with book reviews of the work "Introduction to informetrics. Quantitative methods in Library, Documentation and Information Science" by Egghe and Rousseau. At a first glance the first record (Figure 4; top) looks like a normal record, since the title and authors of the book have been correctly identified. However, the record actually points to a review of the book published in Revista Española de Documentación Científica. The second record (Figure 4; bottom), is also a review of the book which was published in Aslib Proceedings. In this case, the author of the review is the one who appears in the GS record (Brookes).

Figure 4. Authorship and attribution of book reviews

d) Incorrect attribution of documents to authors

Somewhat related to the previous error is the attribution of a document to the wrong authors. In Figure 5 we observe a special case: the book "Introduction to informetrics. Quantitative methods in Library, Documentation and Information Science" by Egghe and Rousseau, is wrongly attributed to Tague-Sutcliffe, probably because this author has a short publication in the journal Information Processing & Management (Figure 5; bottom) with a similar title ("An introduction to informetrics").

Figure 5. Authorship improperly assigned in Google Scholar

e) Failing to merge all versions of a same document into one record

Although the algorithms for grouping versions work well in most cases, Google Scholar sometimes fails to realize that two or more records it has indexed actually represent the same document. This happens when there are enough formal differences between the metadata of the two versions (differences in the way the name of the authors have been stored, in the title, the year of publication…), that Google Scholar judges they're not similar enough to be the same document. This issue mostly affects document types other than journal articles (books, book chapters, reports), but duplicate articles also exist. Articles translated into one or more languages are an extreme example: in those cases, the title of the original version is completely different to that of the translated version, so it is understandable that Google Scholar doesn't realize they are the same document. From a bibliometric perspective, however, their citation counts shouldn't be split.

This issue obviously affects the citation count of some documents. In Figure 6 we can observe how this phenomenon affects a book chapter: "Measuring science", by Van Raan.

Figure 6. Versions of book chapters improperly tied in Google Scholar

f) Grouping different editions of the same book in a single record

Conversely to the previous error, Google Scholar sometimes groups together records that should stay separate, for example in the cases when there are different editions of the same book (a new book edition provides new content, contrary to a reprinting of a book, which is identical to the previous printing).

Figure 7 illustrates the case of "Little Science, big Science", written by Price. This book was first published in 1963 by Columbia University Press, and reedited in 1986 under the title "Little science, big science… and beyond", an edition that contained the original text of the book, as well as seven of his most famous articles.

Figure 7. Different book editions tied in Google Scholar

The primary version (which has received 4,130 citations) is the edition from 1986, but among its versions are several records pointing to the version from 1963. Different editions of the same book should be treated as separate documents when computing citations because their content may be very different. Of course, aumatically detecting and managing these details is a very complex task, and only a very tiny fraction of the documents indexed in Google Scholar (the most influential manuals and seminal works) would benefit from this thorough treatment. We must not forget that Google Scholar is, first of all, a search tool devoted to helping researchers find academic information. A great percentage of users probably don't care about the different editions of a book, and those who do probably just want the most recent one. That may be the reason why Google Scholar usually displays the most recent edition of a book as the primary version. The use of separate entries for different editions is something just a few people, like librarians, would be interested in. In any case, this may have an important effect in citation counts because citations to different editions (providing different content) are added together. In Figure 8 we can see how the 1986 edition of the book is receiving citations that were actually made to the original work published in 1963.

Figure 8. Citations to different book editions tied in Google Scholar

g) Improper attribution of citations to a document

Document citation counts in Google Scholar are also affected by the attribution of "ghost" citations to documents, that is, citations that aren't actually there when we examine the citing document. Figure 9 shows an example of this issue: the work "Le transfert de l'information scientifique et technique: le role des nouvelles technologies de l'information face à la crise du modèle actuel de communication écrite" has allegedly received eight citations, but if we manually examine the second document in the list (marked in red), we can't find any mention of the cited work. This phenomenon has been frequently observed in documents stored in the E-LIS repository.

Figure 9. Appearance of false citations

h) Duplicate citations

This phenomenon is a consequence of an issue previously discussed. When Google Scholar fails to realise that two records are actually versions of the same document, these versions are stored as if they were different documents. Therefore, each of them provides its own set of citations to the citation pool. Since the two sets of citations are probably identical, each cited document will receive two citations from what is actually only one document, thus falsely inflating their citation counts. In Figure 10 we observe a double example of this phenomenon. In the first case (first red rectangle), there are three versions of the same document. Note the differences in the way the authors' names are stored, since this is probably the reason why the records weren't merged into one. In the second case (second red rectangle), the two records refer to the same document (the first one is the English version of the article, and the second one is the Spanish version).

Figure 10. Duplicate citations in Google Scholar

i) Missing citations

There are cases when Google Scholar's parser fails to match a cited reference inside document, with the record of the document it is citing. When Google Scholar parses the reference section within an article, it tries to find a match for these references in its records, but if for some reason the reference hasn't been correctly recorded (authors of the citing article may have made a mistake when citing it or used an uncommon reference format Google Scholar doesn't understand) the system will be unable to make the connection between the two documents. However, we also find examples in which no apparent mistake has been made in the citing document, but still the citation isn't attributed to the cited document. In order to illustrate this issue, in Figure 11 we show how a document ("How to cook the university rankings") is citing in its reference section other document (a doctoral thesis). However, this citation doesn't appear as one of the 13 citations that the thesis has received according to Google Scholar. The reason is unknown. At the time the citing

document was first indexed, the connection wasn't made for some reason, and this error hasn't been solved since. Typos in the PDF can also generate this kind of error.

Figure 11. Citations unrevealed in Google Scholar

2. Mistakes identified in the elaboration of bibliographic profiles

All the errors previously described are related directly with the Google Scholar database (and are concerned with how the automatic parser works). Next we show some of the mistakes identified in the elaboration of bibliographic profiles through Google Scholar Citations:

a) Duplicate profiles Since the only restriction to create a public academic profile in Google Scholar Citations is to provide a valid email, an author (or anyone really) may create as many profiles as he/she wants. This opens the door to the existence of duplicate profiles, that is, different profiles about the same person. In Figure 12 we present some examples of duplicate profiles of authors related to the field of Bibliometrics. The differences in citation counts between profiles are sometimes quite high (for example, one of the profiles belonging to Ruiz-Castillo achieves 1,843 citations whereas in the second profile the figure goes up to 2,430).

Figure 12. Duplicate profiles in Google Scholar Citations
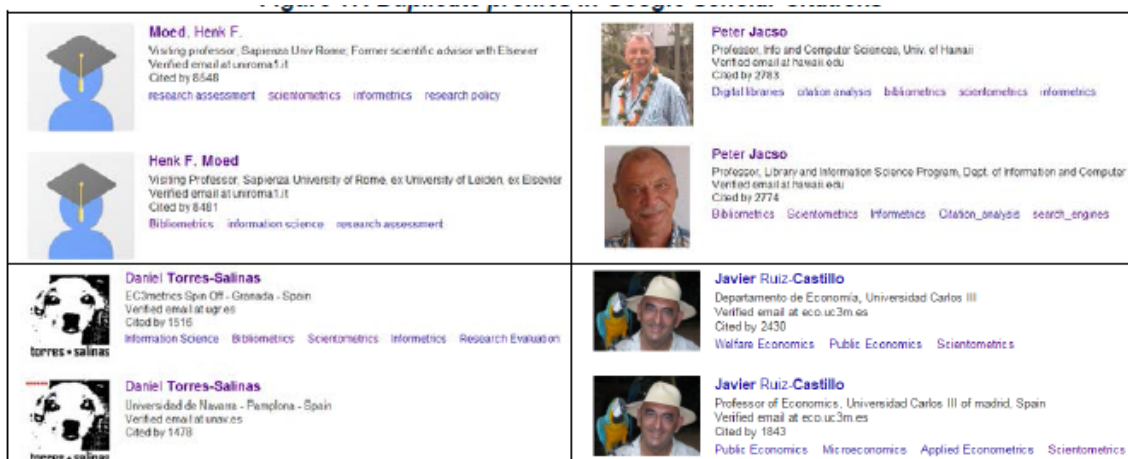
A real problem can arise when one of the profiles has been created by someone other than the author the profile is about. The author may send a request to Google Scholar to delete the profile, but this kind of requests might take a while to be processed, generating a feeling of helplessness in the author.

b) Variety of document types (including non-academic documents), one of the main criticisms to the profiles in Google Scholar Citations (when considering whether they're suited for evaluation purposes) is the inclusion of a wide variety of document types: from peer-reviewed articles to posters. An author can add any kind of work to his profile, and sometimes they aren't

even academic works: teaching materials, software, online resources, etc. (Figure 13). While this is a true shortcoming from the research evaluation perspective, these profiles are designed to showcase any material that the author considers appropriate, especially if these materials could potentially generate some kind of impact through citations. The possibility to select the document typology (as ResearchGate does) may help solve this problem. However, the selection of document type is only an internal mechanism not reflected in the public profile.

Figure 13. Teaching materials in Google Scholar Citations

c) Inclusion of missattributed documents in the profile

The Google Scholar team doesn't oversee the validity of all the information available in Google Scholar Citations. Therefore, it is the sole responsibility of the author that the information visible in his/her profile is accurate. Profiles can be set to be updated automatically (when the system finds an article that it's reasonably sure it's yours, it is automatically added to your profile), or by asking the author for confirmation first when the system thinks an addition or a change should be made. If the user selects the automatic updates, there is a risk that the system will add documents to the profile that the author hasn't actually written, thus falsely increasing the author's bibliometric indicators. The author will probably be completely oblivious to this issue if he or she doesn't check the profile regularly. If that is the case, it shouldn't be considered an active attempt to fake one's bibliometric indicators, but it is still a matter that should be fixed as soon as it comes to the author's knowledge. In Figure 14 we can see an example: the third document (marked in red), which has received 40 citations, hasn't been written by the owner of the profile (Imma Subirats-Coll).

Figure 14. Misattributed documents in Google Scholar Citations

We can find examples where the owner of the profile has participated as a translator or editor of a work (Figure 25). The assignation of the citation counts of a work to the people who have

fulfilled this kind of roles is controversial. At the very least, they should make sure that their role is clearly stated and visible in the profile.

Figure 15. Edition and translation roles in Google Scholar Citations

d) Deliberate manipulation of documents and citations in Google Scholar

Another issue is that of the conscious manipulation of profiles by their owners. The fact that anyone, without advanced technical skills, can manipulate his/her own bibliometric indicators, or other people's (Delgado López-Cózar, Robinson-García & Torres-Salinas, 2014) may affect the credibility of GSC academic profiles if no action to control this issue is taken by the Google Scholar team. In Figure 16 we observe how uploading a set of fake documents to a repository (with nonsensical text, and a list of references which include the set of documents whose impact you want to boost) will, in just a few days, cause the desired adulteration of citation scores in the profiles of the authors of the referenced documents.

Figure 16. Effect of data manipulation in Google Scholar Citations

e) Duplicate documents in profiles

This is also a side effect of the cases when Google Scholar fails to group together different versions of the same document. The consequence for the profiles is that the different versions will also be added as different records in the profile, which might affect (positively or negatively) indicators like the hindex and the i-index, which are computed automatically. Fortunately, profile users can manually merge records in their profile, which will solve this issue (Figure 17). This merge only affects the author's profile. It doesn't alter Google Scholar search query results in any way, that is, there will still be two (or more) records for that document in Google Scholar's index, at least until the error gets fixed in a future update.

Figure 17. Versions not tied in Google Scholar Citations

f) Unclean document titles

This error is also inherited from Google Scholar's metadata parsing errors. Google Scholar Citations allows authors to modify almost all aspects of a record in their profile, including the title of the documents. Unfortunately, not all authors pay attention to such details, and so these errors persist (Figure 18).

Figure 18. Parse errors in identifying document titles in Google Scholar Citations

g) Missing or uncommon areas of interest

One last limitation that may affect the results of this Working Paper is related to the areas of interest declared by the authors in their profiles (a maximum of five areas can be provided). Researchers in bibliometrics with a public profile in Google Scholar Citations, but haven't declared any area of interest (Figure 19, top), those who use uncommon keywords, or keywords in a language other than English may have been overlooked.

Figure 19. Missing (top) and uncommon (bottom) areas of interest in Google Scholar Citations

NOTA: este texto es una adaptación de:

- Martin-Martin, A., Orduna-Malea, E., Ayllon, J. M., & Lopez-Cozar, E. D. (2016). The counting house: measuring those who count. Presence of Bibliometrics, Scientometrics, Informetrics, Webometrics and Altmetrics in the Google Scholar Citations, ResearcherID, ResearchGate, Mendeley & Twitter. arXiv preprint arXiv:1602.02412.

Unidad de Bibliometría

Vicerrectorado de Investigación, Desarrollo e Innovación

Teléfono: 928 451 030

Fax: 928 459 699

Correo:vidi_bibliometria@ulpgc.es

Ubicación: Sede Institucional