

# Productivity and Impact Analysis of a Research and Technology Development Center using Google Scholar Information

Eric Garcia-Cano<sup>a</sup>, Eugenio López-Ortega<sup>b</sup> and Luis Alvarez-Icaza<sup>c</sup>

*Instituto de Ingeniería, Universidad Nacional Autónoma de México, Ciudad Universitaria, Ciudad de México, Mexico*

**Keywords:** Automated Bibliometric Analysis, Strategic Planning, Scientific Research Productivity Metric, Google Scholar.

**Abstract:** This paper presents a project aimed at evaluating the scientific productivity and impact of a Mexican research and technological development center. The proposed evaluation is based on an automated bibliometric analysis system that exploits Google Scholar (GS) information. The import process of the evaluation is shown, including different aspects such as information request times, data verification through a parallel query in Crossref and homogenization of publication sources. As a result, 8,492 documents by 137 researchers associated with the research center were identified. These documents have received 74,683 citations. GS includes a great variety of published materials, such as journal papers, books, conference proceedings, white papers, and technical reports. This diversity of documents allows for a broader evaluation that takes into consideration other types of research products that are not usually considered for assessing scientific productivity. From our work, we conclude that the information in GS can be used to conduct a formal analysis of the productivity and impact of a research center.

## 1 INTRODUCTION

The Institute of Engineering (IIUNAM), belonging to the National Autonomous University of Mexico is a Mexican center for research and technological development that serves various areas of Engineering. A bibliometric analysis project aimed at detecting strategic research topics for its development started in 2012. In its initial stages, a computer system was developed to acquire data from Scopus, and to generate reports related to the behavior of specific research lines.

Google Scholar (GS) information coverage is higher than other databases such as Scopus and Web of Science (WoS) (Harzing, 2014, Harzing and Alakangas, 2016, Harzing and Van Der Wal, 2009, Mingers and Lipitakis, 2010). Several authors have estimated that GS captures more than twice as many citations as WoS and Scopus. Also, it considers multiple types of publications, such as books, congress proceedings, white papers, public reports (norms, regulations), and doctoral theses (Halevi et

al., 2017, Meho and Yang, 2007, Meier and Conkling, 2008).

Some of the most cited works carried out by IIUNAM researchers correspond to books and congress proceedings. Thus, in 2017, the Board of Directors of the Institute decided to analyze the impact, productivity and characteristics of its research based on the information contained in Google Scholar.

This paper presents a review of the process for the information retrieval from GS, as well as the main results, obtained related to the scientific research productivity and impact of the IIUNAM.

## 2 GOOGLE SCHOLAR STRUCTURE

Information gathering is one of the hardest tasks in data analysis. Unlike WoS and Scopus, GS does not provide a public API to access its information

<sup>a</sup>  <https://orcid.org/0000-0002-2723-075X>

<sup>b</sup>  <https://orcid.org/0000-0002-5687-8934>

<sup>c</sup>  <https://orcid.org/0000-0001-9516-3950>

automatically. Therefore, we collected the data by parsing the HTML code with a web scraper.

## 2.1 Site Map

In our analysis, we identified three levels of depth in the way Google Scholar displays information in its database.

### 2.1.1 First Level (L1): Profile Search Results Pages

This first level corresponds to the results pages for authors' profiles search based on specific keywords (in this case, the name of the institution). At this level, GS identifies a list of all author profiles that meet the search criteria.

As shown in Figure 1, this page displays the primary data of the authors: (1) name and link to their personal profile page, (2) photograph, (3) affiliation, (4) email address domain, (5) their areas of interest (if provided), and (6) the total count of citations he or she has received.



Figure 1: GS profile search results page.

### 2.1.2 Second Level (L2): Author Profile Pages

This level corresponds to the authors' profile pages, where all their works are listed. Any researcher can create their account and select the works of their authorship. In this way, GS seeks to deal with its lack of rigor in indexing by relying on authors maintaining and updating their works data, correcting errors and combining duplicate records (Aguillo, 2012, Huang and Yuan, 2012).

In this page (Figure 2), we identify four categories of information:

1. **Author Details:** Basic information provided by the authors, including name, affiliation, fields of interest, email domain (verified by Google), and personal homepage.

2. **Citation Metrics:** GS shows metrics for authors estimated based on the citations collected. Additionally, a bar graph helps to visualize how an author's citations are distributed each year.
3. **Published Works:** The list shows the essential information for each work belonging to the author. The information of each work listed includes:
  - Document title and link to details
  - Citations count and link to the list of citations
  - Year of publication
  - List of co-authors
  - Source information, such as title, volume, and issue number
4. **List of Co-authors:** Authors can enter manually the list of co-authors with whom they frequently collaborate.

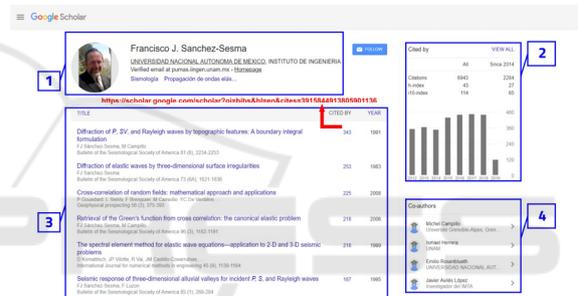


Figure 2: Author Profile Page.

### 2.1.3 Third Level (L3): Work Details View

This last level of depth shows information about each published work. The information consists of a detailed view (Figure 3) of the bibliographic data containing the most comprehensive information GS can provide about a work, such as (1) link to the full text, (2) type of document, (3) editor, (4) abstract, (5) a bar chart with the distribution of citations by year, and (6) the list of versions and related articles. However, it is common that some documents have some missing information. Knowing the type of document is particularly relevant, as in disciplines such as engineering, many high-quality research results are patented and reported in formats other than a journal article or conference paper (Harzing and Van der Wal, 2008, Huang and Yuan, 2012).

Also, if the document has citations reported by WoS, GS shows its citation count as reported in WoS (see Figure 4).

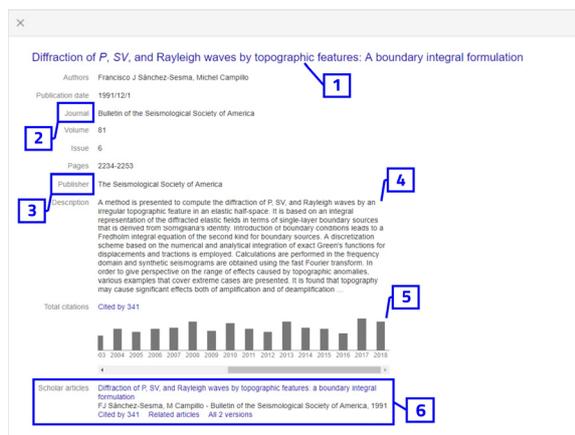


Figure 3: Work details view.

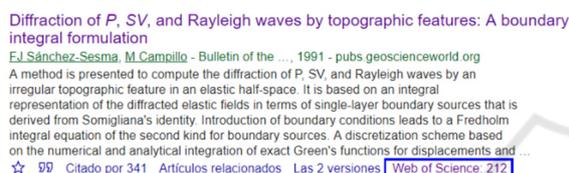


Figure 4: Work citations reported by Web of Science on GS.

The public academic information provided by Google Scholar is vast. In our work, we selected the following data for our analysis:

1. Authors:
  - Name
  - Profile page URL
  - Citations count
  - Citations distribution
2. Work:
  - Title
  - Authors list
  - Details view URL
  - Document type
  - Publication year
  - GS citations count and distribution
  - GS citations page URL
  - WoS citations count and URL
  - Source details

### 3 DATA COLLECTION

To automatize the process to download the information from GS is not an easy task. If too many requests are generated in short time, GS qualifies you as malicious traffic.

We developed an adaptive request rate (ARR) technique, which allows our scraper to collect

information at a moderate speed, without exceeding the established limits per second, minute and hour.

Figure 5 is a diagram of the scraper architecture and data flow in the download process. The main feature of this application is ARR. It is composed of three requests gauges, which act as traffic lights. Each gauge comprises a timer and a request counter. They record the number of queries sent to GS per unit of time in seconds, minutes and hours and adjust the requesting speed to stay within the average set for each unit.

The speed setting is palliative, which means that as many requests are made as possible in a time interval. When the interval is met, it is checked whether the respective limit has been exceeded. When it is exceeded, it calculates the time that the program must wait before it can continue performing requests (compensation time). This time is calculated based on the equation [1], where  $r_{lu}$  are the requests sent in the last unit of time and  $l_{rr}$  is the limit average request rate. Once the pause is over, the scraper restarts the requesting.

$$ct = \left( \frac{r_{lu}}{l_{rr}} \right) - 1 \tag{1}$$

#### 3.1 Data Cross-checking

Every retrieved record from GS was cross-checked with Crossref, which provides a public API to search automatically by work title. This helped us to ensure the reliability of the dataset.

It took an average of 2 to 5 seconds to get an answer from Crossref servers. This active waiting time was included in the pauses generated by the ARR; this helped to slow down the speed of the requests sent to GS.

The categorization of Crossref works is much more detailed; it contains 27 different types. From the types provided by both data sources, we created the common work classification of Table 1.

### 4 RESULTS

In this work, we present indicators and metrics based on descriptive and exploratory analyses of information from GS. This allows us to measure the impact of the productivity research in our institution. Also, this could be of interest for Research and Technological Development Centers, where the results of their research are not always published as articles in peer-reviewed journals.

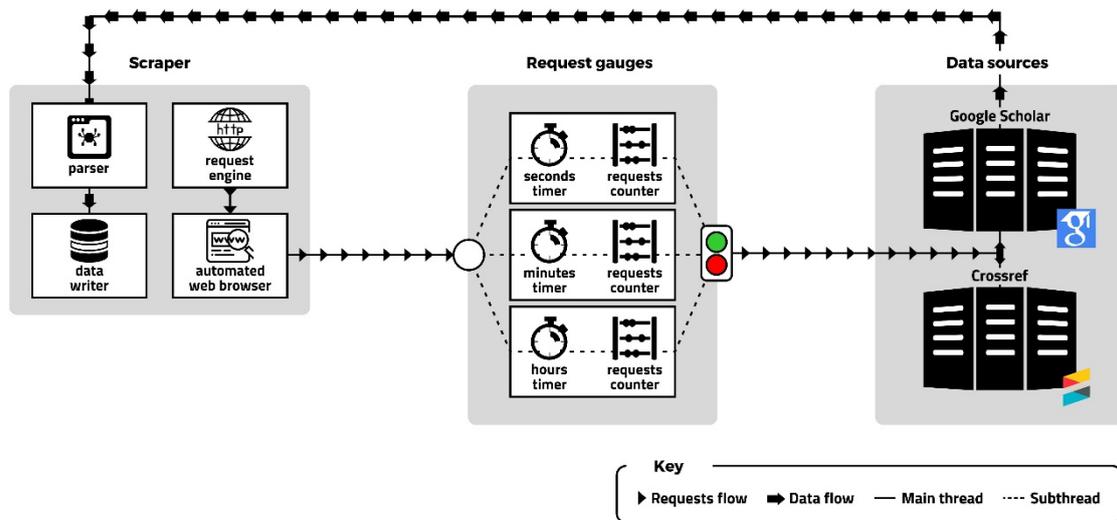


Figure 5: Scraper architecture and download flow.

The "raw" dataset collected from GS is composed of 137 authors associated with our research center with a total of 8,492 works, in 2,565 sources and 74,683 citations. The citations are reduced to 26,972 for the 1,552 works also found in WoS (see section 2.1.3).

Since it is common to have several authors in a paper, it is expected that the document appears in each co-author's profile. When the co-authors are from the same institution, the work will repeatedly appear in the raw dataset. By filter out duplicate works, we retain 6,949 different works. Table 2 presents the metrics used in our analysis.

#### 4.1 Works Analysis

From the data extracted from GS related to the authors' profile, a total of 5,680 (82%) of the works have a GS type. Within these works 3,001 (43%) also have a Crossref type.

From the types provided by both data sources, we create a common classification that divides the works into 7 classes (as seen in Table 1).

Our analysis agrees with other studies; GS has greater work coverage than other databases. In this particular case, less than half of the documents indexed by GS were found in Crossref. From the classified papers, only 1,268 are contained in a JCR journal.

The pace of publication of our research center can be understood by observing academic production and productivity. Production means the number of papers published by year while productivity indicates the average number of documents published per author in each year. Productivity is also used to measure the efficiency of human resources.

Table 1: Common work classification based on Google Scholar and Crossref types.

Common class	Google Scholar Type	Crossref type
Book	Book	book-chapter book-track book-part book book-set reference-book book-series edited-book reference-entry
Journal article	Journal	journal-article journal journal-volume journal-issue
Proceedings article	Conference	proceedings-article proceedings
Thesis	Institution	dissertation
Patent	Patent Number	N/A
Miscellaneous	N/A	monograph report report-series component standard standard-series posted-content (preprints) dataset
Unclassified	Source	other

Table 2: Metrics in the dataset.

Metric	Count
Authors	137
Works	6,949
Sources	2,450
Citations	74,683
WoS citations	26,972
Citations per author	420
Works per author	20.22
Avg. citations per work	10.74
Avg. authors per work	4.47

Figure 6 shows the production of works by year. Figure 7 shows the number of authors who have published each year, along with the average number of papers per author. According to the graph, it is clear that not always more researchers publishing means a higher average number of papers. In 2004, with 68 authors, more works were produced than two years later with 80 authors.

Table 3: Distribution of the works in the dataset.

Class	Works count	Citations count	Average citations per class
Journal article	4 513	60 774	13
Proceedings article	888	4 400	4
Book	284	4 266	15
Thesis	61	398	6
Misc	9	2 839	326
Patent	9	118	13

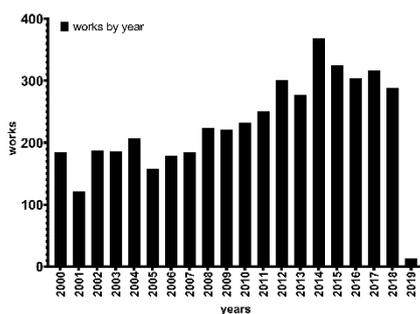


Figure 6: Production of works of our institution by year.

## 4.2 Citations Analysis

Nowadays, the role of bibliometrics for analyzing research impact has increased its relevance. In this sense, citations received for a paper, especially those that are not self-citations, are considered as one of the

most important quality indicators at present, since they speak of peer recognition.

We also were interested in knowing how many of the citations generated by the staff corresponded directly to the research center. Usually, citations reported in a paper are assigned to each author. Works with many citations are not necessarily representative of an individual when there are many co-authors in a single work. Also, it is well known that the higher the number of authors, the more self-citations tend to increase; these situations are known as the co-authorship effect (Batista et al., 2006, Hirsch, 2005).

To minimize the impact of this situation, in our analysis we normalized the citations by work, which means that for each work, the total number of citations is divided by the number of co-authors.

From the normalized citations, we generated an indicator called *institutional citations*. This indicator is calculated as follows. If a paper has 264 citations and 8 authors, and 2 of them are from our institution. First, each author has 33 standard citations ( $264 / 8$ ). Since there are two authors from our institution, then there are 66 institutional citations ( $33 \times 2$ ). Figure 8 shows the total citation versus institutional citations. In average, the institutional citations represents 25% of the total of citations.

Thus, the *participation* that a research center generates in citations is directly related to the number of institutional authors participating in the works.

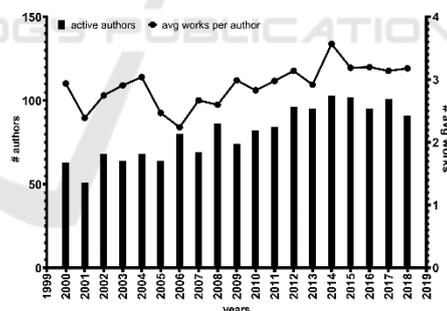


Figure 7: Productivity of our institution.

Another quantitative point of view to study the publication patterns in an institution is to know what is published and where. 80% of the citations are contained in 342 sources representing 34% of the works. In JCR, there are 94 of those sources reporting 32,697 citations in GS.

Taking into account Table 3 and Table 4, it seems evident that the most common work and source types are those classified as *Journal*. It is a fact that its citation and work counts are much higher than any other classification.

The average number of citations by classification is obtained by dividing the total number of citations of one class, by the number of works in the same class, this normalized the data to make comparable all work categories, putting them all on the same scale.

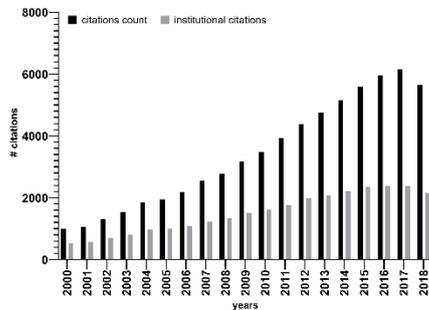


Figure 8: Citations count vs. institutional citations.

Because of the above, we observe that the work class *Misc*, which includes reports of different natures, is the most important within the dataset, followed by the *Book* class, moving *Journal* to third place. This is congruent with the results shown in Table 4, with *Misc* and *Book* publications appearing with a higher weight than *Journal*; also, the two most cited works in the dataset are a report and a book.

Journal articles will continue to be the most numerous type of publication, as citations in indexed journals are the standard measure of the quality of current research and have a direct impact on the professional lives of researchers, such as in recruitment, tenure, promotion or funding. In the case of institutions, they are reflected in university rankings (Martín-Martín et al., 2018a).

### 4.3 WoS Citations Comparison

Google Scholar displays the WoS citation count within its indexed works for subscribers to that database, and we get it as seen in section 2.1.3.

Table 4: Citation coverage by publication type.

Source class	Works count	Citations count	Average citations per class
Journal	1 936	49 396	26
Conference	437	4 809	11
Book	21	912	43
Misc	1	57	57

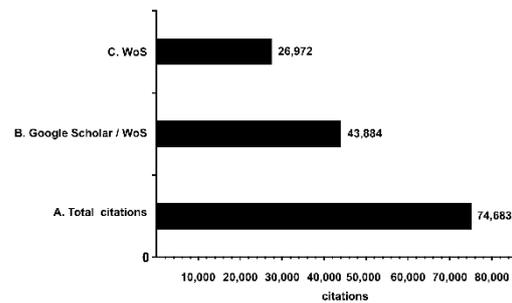


Figure 9: A. Total citations reported by GS for all works. B. Citations reported in Google Scholar for the works that are also reported citations in WoS. C. Citations in WoS for the works in B.

There are several comparative studies about the citation coverage between GS and other databases. In summary, they conclude that the citation count is about three times higher in GS than in WoS and Scopus (Halevi et al., 2017, Harzing, 2013, Mingers and Meyer, 2017).

In the study made by (Martín-Martín et al., 2018b), they performed an analysis of the citations overlap, 2,515 highly cited documents were included from the sources mentioned above. They found that, overall, citations in all three databases matched 46.9% and that more than one-third of the total citations were only found by GS. By disaggregating the results by fields of knowledge, they discovered that nearly all citations overlapped in the areas of *Engineering*, *Physics & Mathematics*, *Earth Sciences*, and *Chemistry & Materials*, with overlap

Table 5: Results obtained from the chi-square test for the distribution of WoS citations.

Paper rank in <i>h</i> index	GS citations count	WoS citations count	Degrees of freedom	X <sup>2</sup>	P value
9	488	332	7	14.33	0.0456
31	194	34	17	29.14	0.0330
39	177	69	5	17.6	0.0035
43	91	46	15	29.89	0.0123
52	54	25	15	29.19	0.0152
56	121	81	17	45.71	0.0002
107	109	61	16	43.23	0.0003
All citations	12 890	7 239	18	90.85	0.0000

percentages ranging from 46.8% to 67.7%. These results lead them to conclude that GS contains almost all reported WoS citations (95%).

Figure 9 shows the results we obtained from our dataset. Considering the institution as a whole, GS identified three times more citations than WoS; it found more documents associated with institutional researchers. By comparing the works identified by GS that are also in WoS, we found a citation overlap of around 60%, which coincides with the range mentioned above for the research areas of our interest.

GS only provides the citation count reported in WoS, but not its distribution. It has been found that there is a remarkably strong linear correlation at the document level between GS, Scopus, and WoS citation counts in the range of 0.8 to 0.9. This degree of correlation suggests that the behavior of citations is similar in all three data sources, regardless of the volume of documents in each of them (Halevi et al., 2017, Martín-Martín et al., 2018a).

Considering the findings reported for overlap and correlation of citations previously mentioned, we generated a theoretical distribution by year of the citations count reported in WoS, based on the GS distribution by year. For this analysis, we consider 73 documents of the institutional  $h$ -index that contain citations in both databases, between 2000 and 2018. We performed the Pearson chi-square test that measures the discrepancy between an observed and a theoretical distribution, using the actual distribution of WoS citations with a statistical significance of 95%.

At the document level, we find that in only 9.5% of cases, the theoretical and observed distributions are adjusted with a minimal discrepancy. When considering the citations of the 73 papers as a unit, there is also a statistically significant difference (Table 5). These results suggest that the proposed distribution model should be applied only to distribute citations at the institutional level (Figure 10).

#### 4.4 Institutional Quality and Impact

Nowadays, the most accepted index to consider when it comes to characterizing a researcher's work is the  $h$ -index. (Hirsch, 2005) defines this index as the  $h$  articles of a researcher with a number of citations equal to or greater than  $h$ . The original research focused only on physicists, but the index has proven been useful for many disciplines.

Hirsch also mentioned several disadvantages of single-number metrics that are commonly used to evaluate scientific research, generally referring to the

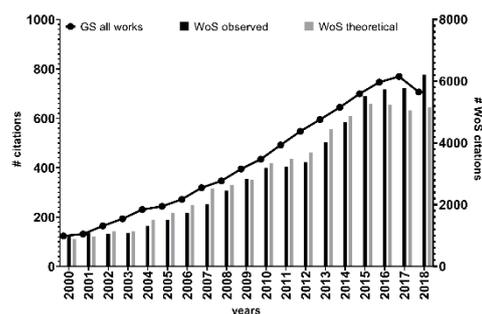


Figure 10: Goodness of fit for the theoretical distribution of WoS citations generated from the GS distribution.

counting of citations and works. His main criticisms are based on the difficulty of the calculation, the low relevance of the impact and the use of arbitrary parameters. Hence, the calculation of the  $h$ -index is necessary to understand the quality of the research conducted at the institutional level.

Following the mechanism indicated by the index creator, we consider all authors' works, ordering them according to their citations, until we find that 108 papers have at least 108 citations. Documents of  $h$ -index are published in 59 sources (24 of them in JCR), with the participation of 33% of the authors, adding a total of 25,808 citations in GS and 8,033 in WoS. As expected, *Journal* is the most extensive work class. However, the ones with the most significant weight are these classified as *Misc* and *Book*. We also calculated a normalized  $h$ -index (106), similar to that proposed by (Batista et al., 2006) considering normalized citations, instead of total citations. The purpose of this index is to determine the impact per author. We also propose calculating an institutional  $h$ -index (65) computed from institutional citations.

## 5 CONCLUSIONS

Today, bibliometric analysis has become a fundamental tool for assessing the scientific research quality, its impact and collaboration patterns, often used in strategic planning and evaluation contexts. Google Scholar is a valuable source of information for these purposes, having clear advantages over WoS and Scopus, in terms of citations coverage and types of work generated in research and technological development centers. In this study, we explore the extent to which the data available in Google Scholar can be used to conduct a formal analysis of an institution's research, concluding:

- GS can be used as a source of bibliometric data for authors and institutions, whose research

results are published in formats that are not necessarily journal articles and that are in a language other than English.

- The grouping of authors through their affiliation and the possibility of listing them by that criterion in GS makes possible the recovery and analysis of an institution's scientific production as a whole.
- GS provides enough information to calculate metrics commonly used to estimate the scientific quality of research and to discover publication patterns from a wide variety of publications.
- The recovery of the information from GS is a challenging task since Google does not provide an API to access its data.
- Consistent with previous literature, we found that GS coverage is greater than other scientific databases.
- As corroboration of our productivity metrics, we observed a significant correlation in citation counts for the publications of GS that overlapped with those of WoS and Scopus.

Finally, our analyses were limited to the scientific publications of authors of our institution. As a future work, we will include other institutions with similar lines of research to compare and evaluate the performance of different research centers.

## REFERENCES

- Aguilo, I. F., 2012, Is Google Scholar useful for bibliometrics? A webometric. *Scientometrics*, 91(2), 343-351.
- Batista, P. D., Campiteli, M. G., & Kinouchi, O., 2006. Is it possible to compare researchers with different scientific interests? *Scientometrics*, 68(1), 179-189.
- Halevi, G., Moed, H., & Bar-Ilan, J., 2017. Suitability of Google Scholar as a source of scientific information and as a source of data for scientific evaluation—Review of the Literature. *Journal of Informetrics*, 11(3), 823-834.
- Harzing, A.-W., 2013. A preliminary test of Google Scholar as a source for citation data: a longitudinal study of Nobel prize winners. *Scientometrics*, 94(3), 1057-1075.
- Harzing, A.-W., 2014. A longitudinal study of Google Scholar coverage between 2012 and 2013. *Scientometrics*, 98(1), 565-575.
- Harzing, A.-W. K., & Van der Wal, R., 2008. Google Scholar as a new source for citation analysis. *Ethics in science and environmental politics*, 8(1), 61-73.
- Harzing, A.-W., & Alakangas, S., 2016. Google Scholar, Scopus and the Web of Science: A longitudinal and cross-disciplinary comparison. *Google Scholar, Scopus and the Web of Science: A longitudinal and cross-disciplinary comparison*, 106(2), 787-804.
- Harzing, A.-W., & Van der Wal, R., 2009. A Google Scholar h-Index for Journals: An Alternative Metric to Measure Journal Impact in Economics and Business. *Journal of the American Society for Information Science and technology*, 60(1), 41-46.
- Hirsch, J. E., 2005. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46), 16569-16572.
- Huang, Z., & Yuan, B., 2012. Mining Google Scholar Citations: An Exploratory Study. *International Conference on Intelligent Computing* (págs. 182-189). Berlin: Springer.
- Martín-Martín, A., Orduna-Malea, E., & López-Cózar, E. D., 2018. Coverage of highly-cited documents in Google Scholar, Web of Science, and Scopus: a multidisciplinary comparison. *Scientometrics*, 116(3), 2175-2188.
- Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & López-Cozar, E. D., 2018. Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, 12(4), 1160-1177.
- 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.
- Meier, J. J., & Conkling, T. W., 2008. Google Scholar's coverage of the engineering literature: An empirical study. *The Journal of Academic Librarianship*, 34(3), 196-201.
- Mingers, J., & Lipitakis, E., 2010. Counting the citations: A comparison of Web of Science and Google Scholar in the field of business and management. *Scientometrics*, 85(2), 613-625.