Citation counts and inclusion of references in seven free-access scholarly databases: a comparative analysis

Lorena Delgado-Quirós (ORCID: 0000-0001-8738-7276) and José Luis Ortega (ORCID: 0000-0001-9857-1511)

Institute for Advanced Social Studies (IESA-CSIC), Córdoba, Spain
Joint Research Unit Knowledge Transfer and Innovation, (UCO-CSIC), Córdoba, Spain
ldelgado@iesa.csic.es, jortega@iesa.csic.es

Abstract

The aim of this study is to examine disparities in citation counts amongst scholarly databases and the reasons that contribute to these differences. A random Crossref sample of more than 115k DOIs was selected and subsequently searched across six databases (Dimensions, Google Scholar, Microsoft Academic, Scilit, Semantic Scholar and The Lens). In July 2021, citation counts and lists of references were extracted from each database for comparative processing and analysis. The findings indicate that publications in Crossref-based databases (Crossref, Dimensions, Scilit and The Lens) have similar citation counts, while documents in search engines (Google Scholar, Microsoft Academic and Semantic Scholar) have a higher number of citations due to a greater coverage of publications, but also to the integration of web copies. Analysis of references has revealed that Scilit only extracts references with Digital Object Identifiers (DOI) and that Semantic Scholar causes significant problems when it adds references from external web versions. Ultimately, the study has shown that all the databases struggle to index references from books and book chapters, which may be attributable to certain academic publishers. The study concludes with a discussion of the potential effects on research evaluation that may arise from this lack of citations.

Keywords

Citation counts; Reference processing; Academic search engines; Third-party databases; Coverage analysis

1. Introduction

“Standing on the shoulders of giants” is a popular phrase that alludes to the cumulative nature of science, according to which every new advance is built upon earlier discoveries. In this sense, the validity of scientific thought lies in the ability of new theories and statements to fit or contradict previously consolidated knowledge. The widespread use of citations in research papers evidences the need for scientific research to put findings into context as regards previous analyses. Citations then acquire special importance in scientific activity since they enable research papers to be bound together (thereby enhancing information retrieval) and add value to the impact of specific studies within the scholarly community (thereby bolstering research assessment) (MacRoberts & MacRoberts, 1989).

Unsurprisingly, considerable efforts are made for scientific bibliographic databases to extract and index references, since they facilitate building citation indexes to enlarge search strategies, as well as to design metrics that value and rank academic entities (Garfield, 1955). Originally, subscriptions to journals and agreements with publishers were used by established databases such as Science Citation Index (now Web of Science, WoS) and Scopus to gather references from articles. This procedure is more accurate because references come directly from the source. However, it implies high time and economic costs, because references from different sources and formats require processing, thereby considerably limiting the size of the database. In this sense, these first-generation tools adopt a selective approach, by focusing on covering the core of scientific literature.

The advent of the Web put an end to the practice of indexing only a portion of the most important literature. The increasing availability of research publications on the Web, the digital transformation of the academic publishing model and advances in information processing and storage technologies fostered the appearance of academic search engines (Ortega, 2014). CiteSeer (1997), Google Scholar (2004) and Microsoft Academic (2009) were the first to crawl the Web, searching for scholarly publications regardless of source or typology. This ability allows them more thorough coverage, surpassing traditional scholarly databases in terms of size and variety. Another original feature is that they automatically parse bibliographic references from full text documents, enabling them to compute citations and design bibliometric indicators. However, the main drawback of these platforms is that their primary sources (i.e., webpages) display incomplete and unstructured information about publications, resulting in poor metadata quality. This limitation also affects references parsing, causing the loss of possible citations.

The availability of databases that make open citation information accessible (e.g., PubMed, DOAJ, Microsoft Academic) and the free release of bibliographic references by academic publishers (e.g., Crossref) are currently encouraging the emergence of a new generation of third-party scholarly databases that are fed from other external and open sources. Products such as Dimensions, The Lens or OpenAlex face the challenge of integrating different data formats as well as of processing scientific entities linked to publications (i.e., author disambiguation, identifier assignation and discipline classification). These issues might also affect the citation count when references are incomplete or duplicated.

Due to the high relevance of bibliometric indicators in research evaluation, and the existence of different scholarly databases containing citation information, it is important to explore how bibliographic references are extracted and processed and the extent to which this treatment would bias citation counts and other associated metrics. These results would have important implications for the use and selection of these products in research evaluation.

2. Literature review

Citation extraction and indexation in scholarly databases have been widely studied since they first began. Initially, the coverage of the Science Citation Index, the only generalist citation index, was examined to reveal citation analysis limitations. Carpenter and Narin (1981) demonstrated that this database showed an incomplete coverage of non-English-speaking journals. Nederhof (1985) confirmed that there was also a disciplinary bias in sociologists' citation counts. More recently, Gallagher and Barnaby (1998) showed how this journal coverage bias underestimated the impact factor of peripheral disciplines. Some of these limitations are due to differences between disciplines (Narin, 1976) and document types (Line, 1979) in the use of bibliographic references.

The appearance of new competitors such as Scopus and Google Scholar has prompted more research comparing citation counts across databases. According to Scopus, many papers show that this new database has slightly more citations than the WoS because it contains more journals (Ball & Tunger, 2006; López-Illescas et al., 2008; Mongeon & Paul-Hus, 2016) and is well-balanced by discipline and language (Archambault et al., 2009; Vera-Baceta et al., 2019). However, Google Scholar considerably surpasses both databases in the number of citations, mainly because it captures citations from non-journal documents (Bauer & Bakkalbasi, 2005; Kousha & Thelwall, 2008; Martín-Martín et al., 2018). This extensive coverage leads to critiques about poor inclusion criteria (Yang & Meho, 2006) and duplicate citation counts (Jacsó, 2008).

The appearance of academic search engines (e.g., Google Scholar, Microsoft Academic and Semantic Scholar) attracted the attention of different studies that aimed to explore their potential for research evaluation. When Herrmannova and Knoth (2016) examined the entire Microsoft Academic Graph, they discovered that 76% of papers lacked references, which had a negative effect on the number of citations. Hug et al. (2017) observed that, although Microsoft Academic indexed the same number of citations as Scopus, reference metadata quality was lower. Hannousse (2021) tested Semantic Scholar in relation to Google Scholar, finding that both search engines covered secondary literature similarly. This result was also confirmed by Kacperski et al. (2023) when they compared different confirmation biased queries against Google Scholar and Semantic Scholar.

The recent rise of so-called third-party databases has fostered the comparative study of these products with the purpose of observing their principal advantages and limitations. These databases are characterized by their being fed by external open sources. The first multiple analysis was performed by Harzing (2019), who compared her personal production in Crossref, Dimensions, Google Scholar, Microsoft Academic, Scopus and the WoS. She found that when new databases are compared to Scopus and the WoS, they have comparable or greater coverage of citations, but substantively fewer citations than academic search engines. Visser et al. (2021) also compared the same databases, with the exception of Google Scholar, examining variations in the coverage of the number of references and citations. Their findings demonstrated that documents with a low number of references and citations are underrepresented in Scopus and the WoS. Guerrero-Bote et al. (2021) matched the entire databases of Scopus and Dimensions and pronounced that Scopus had a higher volume of citations than Dimensions.

Martín-Martín et al. (2021) explored in greater detail citation variations between the same platforms, including Google Scholar and COCI, the OpenCitations Index of Crossref. According to their results, Microsoft Academic and Dimensions offered at least as many citations as Scopus and the WoS, and Google Scholar discovered 26% more distinct citations than the other sites. In a more recent study, Delgado-Quirós and Ortega (2024) compared publication metadata of eight scholarly databases. Their findings indicated that third-party databases contain more accurate descriptive information than academic search engines, and that books and book chapters are poorly identified on all platforms.

Despite these efforts to analyze citations across databases, a systematic comparison of citation counts between academic search engines and third-party databases has yet to be done, particularly with regard to the potential impact that bibliographic reference coverage may have on these differences.

3. Objectives

The primary goal of this study is to evaluate and compare the coverage of citations across seven scholarly databases, to identify possible biases and find potential causes. This paper will examine the citation data accessible in each database, the differences between them and the ways in which reference indexation may account for these biases. Three research questions were formulated:

- Which database shows most indicators and offers most citation information?
- Are there notable differences in the citation count between databases, and if so, why?
- How does the handling of references affect the computation of citations? What potential biases might influence which references are included in each database?

4. Methods

4.1. Source selection criteria

This comparative approach entails the selection of the same sample in each database, with the aim of benchmarking citation counts and reference lists for the same publications. Seven bibliographic databases were considered for the study: Crossref, Dimensions, Google Scholar, Microsoft Academic, Scilit, Semantic Scholar and The Lens. Two criteria were considered when choosing these sources:

- Being publicly available on the Internet, with a free-subscription search interface.
- Offering metrics for evaluating the research, or at the very least, citation counts.

4.2. Sample selection and extraction

Multiple reasons explain why Crossref was chosen as the control group. The first was due to an operative cause. The most extensive persistent identifier for research articles in the publishing system, the Digital Object Identifier (DOI), is assigned by the publishers' consortium Crossref. Though it is restricted to publisher members (Visser et al., 2019), its usage is justified because all of these platforms enable publications to be queried by DOIs, which expedites and improves matching.

The second justification stems from methodological concerns: Crossref allows random sample extraction from documents (https://api.crossref.org/works?sample=100). As a result, the sample's representativeness is strengthened because it is not influenced by matching techniques, filters or ranking algorithms that could skew its quality. A third motive is that publishers can request a DOI for any published material, regardless of typology, subject or language. As a result, the Crossref database contains no inclusion criteria that can restrict the coverage of particular document types (e.g., indexes, acknowledgements or front covers). The inclusion policies of the various bibliographic systems could be readily understood if this non-selective criterion were applied.

Finally, Crossref is fed by publishers when they deposit their publications' metadata. This database could therefore be considered the most authoritative source because publishers are assumed to provide the most reliable and accurate information on their own publications.

4.3. Sources description

This section provides some facts about the origin, functioning and coverage of each source analyzed:

- Dimensions: Developed by Digital Science in 2018, it is primarily supported by external products, including Crossref (Hook et al., 2018). In addition to patents (154 million),

datasets (12 million) and grants (7 million), it compiles more than 138 million publications.

- Google Scholar: Obtaining data directly from the Web, this search engine is one of the most significant academic search engines because of its estimated size (389 million) and age (2004) (Gusenbauer, 2019). It can extract citations and content information from paywalled journals thanks to special agreements with publishers. Additionally, books (Google Books) and patents (Google Patents) can be accessed using its search interface.
- Microsoft Academic: This search engine's most recent version ran from 2016 to 2021. It reached 260 million papers by crawling the Web and gathering metadata from scientific publications, in the same way as Google Scholar. A portion of its database (Microsoft Academic Graph) was made publicly available, allowing other bibliographic tools to utilize it.
- Scilit: Developed by publisher MDPI to compete in the scholarly database industry, this database has been in operation since 2014. It indexes 159 million academic papers, primarily from PubMed and Crossref.
- Semantic Scholar: The Allen Institute for Artificial Intelligence introduced this search engine in 2015. Despite using crawlers to gather data from the Internet, it agreed to use Microsoft Academic Graph as a primary source in 2018 (Boyle, 2018). It currently contains almost 214 million academic publications.
- The Lens: This database was created in 2000 by the non-profit company Cambia. Originally a patent database, in 2018, it added academic articles from Crossref, PubMed and Microsoft Academic. More than 225 million scholarly works are currently available.

4.4. Data retrieving

In August 2020, a random sample of 115,885 DOIs was taken from Crossref; the sole requirement was that the papers had to have been published between 2014 and 2018. This time frame was chosen to allow articles to obtain a substantial number of citations. To generate the sample, 1,200 automated requests were made to https://api.crossref.org/works?sample=100. To arrive at the final list, duplicate records generated by this random method were eliminated. The distribution of documents obtained is consistent with the total database (Hendricks et al., 2020), which strengthens the sample's reliability.

Subsequently, queries were made on every platform for this control sample in order to compare the records and retrieve all the publication-specific data. All tasks were completed by July 2021, including recapturing the Crossref sample to obtain citations at the same time as the other databases, thereby resulting in comparable samples.

Each platform's extraction procedure is explained in more detail:

- Dimensions: The following URL (https://app.dimensions.ai/dsl/v2) provided access to this database. We extracted the data using the dimensionsR package in R. The results were downloaded in JSON format because dimensionsR produced issues with the conversion of JSON outputs to the CSV format.
- Google Scholar: Since Google Scholar does not make its data accessible, web scraping was utilized to automatically query each DOI in the search box. To simulate a browser session and avoid anti-robot measures (i.e., captchas), the RSelenium R package was used. The results were completed through a title search using the query

"allintitle:title" because some DOIs could not be indexed (Martín-Martín et al., 2018). Fuzzy Lookup add-in for Excel was used to check the title search, and only records with a similarity with Crossref titles higher than 75% were selected, resulting in 3,828 (4.1%) records.

- Microsoft Academic: Multiple approaches were followed to obtain the sample from this service. First, publications were extracted using DOIs utilizing SPARQL (https://makg.org/sparql) and REST API (https://api.labs.cognitive.microsoft.com/academic/v1.0/evaluate) endpoints. The R package microdemic was utilized to query the API. Nevertheless, we were compelled to download the entire Zenodo table of publications (https://zenodo.org/record/2628216) and use DOIs and titles to locally match the sample because the of low indexation of DOIs (37.1%) and their case sensitivity.
- Scilit: The public API for this platform is https://app.scilit.net/api/v1/. A Python script was created to obtain the data because access can only be granted through the POST protocol.
- Semantic Scholar: https://api.semanticscholar.org/v1 is the public API available for this database. Data extraction was done using the semscholar R program. However, the API was directly queried subsequently, to detect any problems in the retrieval process. A script was written in Python.
- The Lens: This service offered temporary access to its API (https://api.lens.org/scholarly/search) upon formal request. In this instance, the data was extracted directly using a R script. However, in July 2021, the references field could not be correctly obtained for technical reasons. To compensate for this restriction, in January 2023 a small sample of 5,000 records was taken from the original 2021 Crossref sample. This set of publications was queried using the main search page (https://www.lens.org/lens/) to obtain the full list of references, resulting in 4,996 records being successfully obtained in response to the request.

Table 1 summarizes the total number of retrieved publications and coverage percentage according to the Crossref sample. The figures show that each resulting sample is sufficiently wide as to be representative in a cross-database comparison.

Database	Dimensions	Google Scholar	Microsoft Academic	Scilit	Semantic Scholar	The Lens
Dimensions	103,903 (89.7%)					
Google Scholar	94,215 (81.3%)	100,722 (86.9%)				
Microsoft Academic	86,385 (74.5%)	86,403 (74.6%)	91,309 (78.8%)			
Scilit	103,722 (89.5%)	100,187 (86.5%)	91,202 (78.7%)	115,119 (99.3%)		
Semantic Scholar	80,499 (69.5%)	79,938 (69.0%)	76,111 (65.7%)	84,935 (73.3%)	85,201 (73.5%)	
The Lens	102,916 (88.8%)	99,424 (85.8%)	90,479 (78.1%)	114,018 (98.4%)	84,137 (72.6%)	114,307 (98.6%)

Table 1. Matrix including number and percentage of publications retrieved from each database according to the Crossref sample.

5. Results

5.1. Metrics

Indicators are an essential component for classifying and ranking publications in databases. In essence, they all offer the number of citations that each publication receives in the database. However, some databases compute more complex metrics regarding publications. Dimension is the database that offers the greatest number of metrics: Recent Citation (citation received within the last two years), Field Citation Ratio (FCR) (total number of citations by the average number of citations of the field in the same year), Relative Citation Ratio (RCR) (total number of citations by the average number of citations in their co-citation network) and the Altmetric score (Composed altmetric indicator computed by Altmetric.com). Semantic Scholar offers citationVelocity (now depreciated, it was the citation average in the last three years) and influentialCitationCount, a measure based on machine-learning algorithms that assesses the significance of the cited work in the findings of the citing paper. Real (CitationCount) and estimated (EstimatedCitations) citations were the only measures provided by Microsoft Academic.

Some databases not only provide citation counts, but also the list of cited documents, which could be used for more complex bibliometric analyses (e.g., co-citation analysis and bibliographic coupling) (Table 2). Only The Lens (scholarly_citations) and Semantic Scholar (citations) make this data available. However, a references list is supplied by almost every platform, except Google Scholar.

Database	Citations list	Citations count	References list	References count
Crossref	No	Yes	Yes (DOI)	Yes
Dimensions	No	Yes	Yes (ID, DOI)	Yes
Google Scholar	No	Yes	No	No
MAG	No	Yes	Yes (ID)	Yes
Scilit	No	Yes	Yes (ID)	Yes
Semantic Scholar	Yes (DOI)	Yes	Yes (ID, DOI)	Yes
The Lens	Yes (ID)	Yes	Yes (ID)	Yes

Table 2. Information available on references and citations in each database.

5.2. Citations analysis

A first exploratory analysis compares the average variation of citations between databases. This displays the proportion in which a database captures and computes more citations than another, giving an approximation of the possible methods used to calculate citations in each database. These variations are calculated according to the following formula:

$$ACV_{(a,b)} = \frac{\sum C_a - C_b}{(A \cap B)}$$

where the average citation variation ($ACV_{(a,b)}$) is the summation of the number of citations of a set of publications in the database A (C_a) minus the number of citations of that same group of publications in the database B (C_b), divided by the number of publications included in A and B ($A \cap B$) databases. The results of this formula make it possible to indicate positive or negative biases in the coverage or computation of citations of one database according to another.

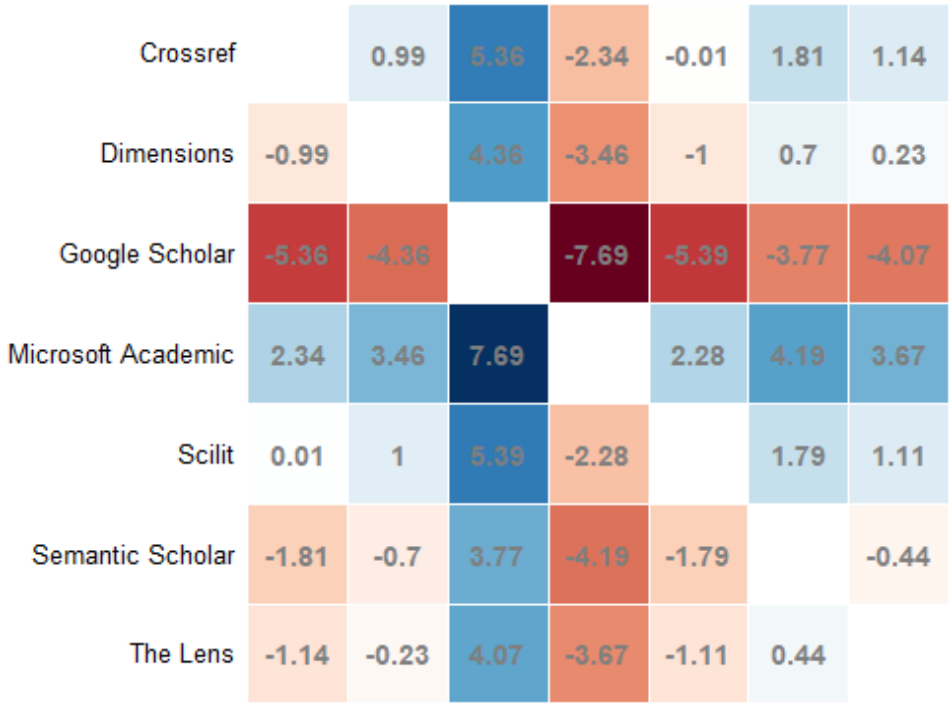

Figure 1. Matrix with the average citation variation among scholarly databases (all pairwise comparisons are significant at p-value<0.001, Wilcoxon rank-sum test).

Figure 1 depicts a matrix showing the average citation discrepancies between databases. The Wilcoxon rank-sum test for paired samples was used to test differences between databases and confirming that all the samples come from different populations (p-value<0.001). This variation enables the strength and direction (positive or negative) of those differences to be determined more accurately than a correlation coefficient. Additionally, their capacity to extract citations can be assessed in relation to the size of the variation. Thus, for example, Crossref's publications have 2.34 times as many citations on average as those in Microsoft Academic, whereas publications in Google Scholar have 5.36 times as many citations on average as those in Crossref. In general, Crossref-based databases (Crossref, Dimensions, Scilit and The Lens) show few variances, ranging from 1.14 for The Lens, according to Crossref, to 0.01 for Scilit, with regard to Crossref.

Nonetheless, there are significant differences amongst academic search engines, which confirms their use of different methods to gather and process citations. Google Scholar captures the most citations, on average 3.77 more citations than Semantic Scholar and 7.69 more than Microsoft Academic. The magnitude of Google Scholar—which is regarded as the greatest scholarly database (Martín-Martín et al., 2018; Gusenbauer, 2019)—contributes significantly to its advantage. But this advantage could also be due to incorrect data integration. The grouping of various editions of the same work could be one significant example. After choosing the papers with more than 200 citations from the Google Scholar sample (N=679), we discovered that 36.2% of the articles had citations aggregated from different editions or versions (books, pre-prints, reprints, etc.) (Martín-Martín et al., 2017).

Conversely, Microsoft Academic has the fewest differences, capturing 2.28 citations fewer than Scilit and 7.69 citations fewer than Google Scholar. The situation of Microsoft Academic is noteworthy due to the relatively low number of citations compared to the size of the database (204M), which is significantly higher than Crossref (125M), Dimensions (130M) or Scilit (149M),

as of August 2022. The cause appears to be a generalized failure to update the database. Taking the Microsoft Academic Knowledge Graph (makg.org) as a reference, we queried publications by publication date from 2016 to 2020. The results showed that there was a generalized fall since 2016 of publications with publication date field filled out: 7.1M in 2016, 6.3M in 2017, 5.2M in 2018, 3.9M in 2019 and 1M in 2020. This trend is supported by Scheidsteger and Haunschild (2023) who also observed that the number of publications decreased from 2017. This slowdown in the aggregation of new records might be the cause of fewer citations to Microsoft Academic publications.

These results demonstrate that the quantity of publications does not always correlate with the amount of citations in a database; rather, citations are the outcome of significant processing tasks that are addressed to compute citations. Examining the inclusion of references in each database can help explain how these variations occur.

5.3. References analysis

In the same way as for citations, an initial approach to differences in the indexation of references is to compare how the number of references vary between different databases. We have employed the same variation formula, replacing citations by references.

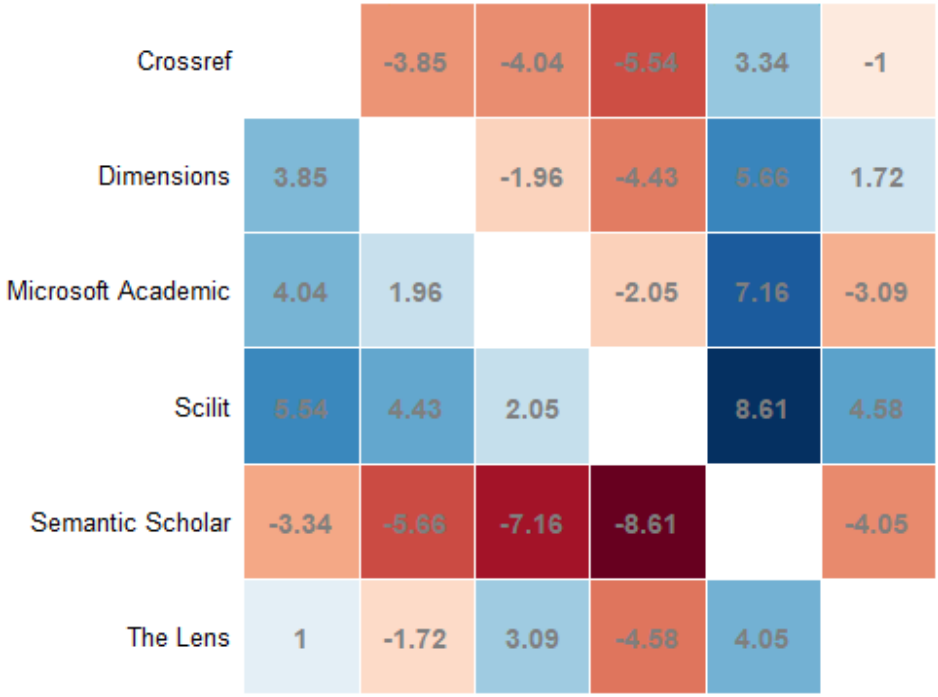

Figure 2. Matrix with the average variation of reference counts among scholarly databases (all pairwise comparisons are significant at p-value<0.001, Wilcoxon rank-sum test).

Figure 2 shows the average differences in the number of indexed references by publication across databases. Google Scholar is excluded from the analysis because the references cannot be retrieved. The Wilcoxon rank-sum test for paired samples was again used to test differences between databases and showed that all the samples are different among them (p-value<0.001). Differences are more noticeable in references, which denotes that the processing of this material differs according to database. In the case of Crossref, all of the references included in the papers are indexed. However, this coverage depends on the publishers' capacity to deposit the entire list of references. Just 59.2% of the papers in the

Crossref sample had references at the time the sample was gathered. Crossref indexes, on average, 5.54 more references than Scilit or one more than The Lens, even when papers with no deposited references are included. These differences are explained because the scholarly databases only index references that point to documents previously included, which is confirmed when all the references in each database point to internal IDs.

Conversely, Scilit is the database that handles fewer references: 2.05 fewer than Semantic Scholar and 8.61 fewer than Microsoft Academic. A possible explanation could be that Scilit only indexes references with a DOI. For example, Scilit only indexed references with DOI in 61.6% of the 31,522 papers with references in Crossref; the other references would therefore have to be obtained from other sources (PubMed, DOAJ, etc.).

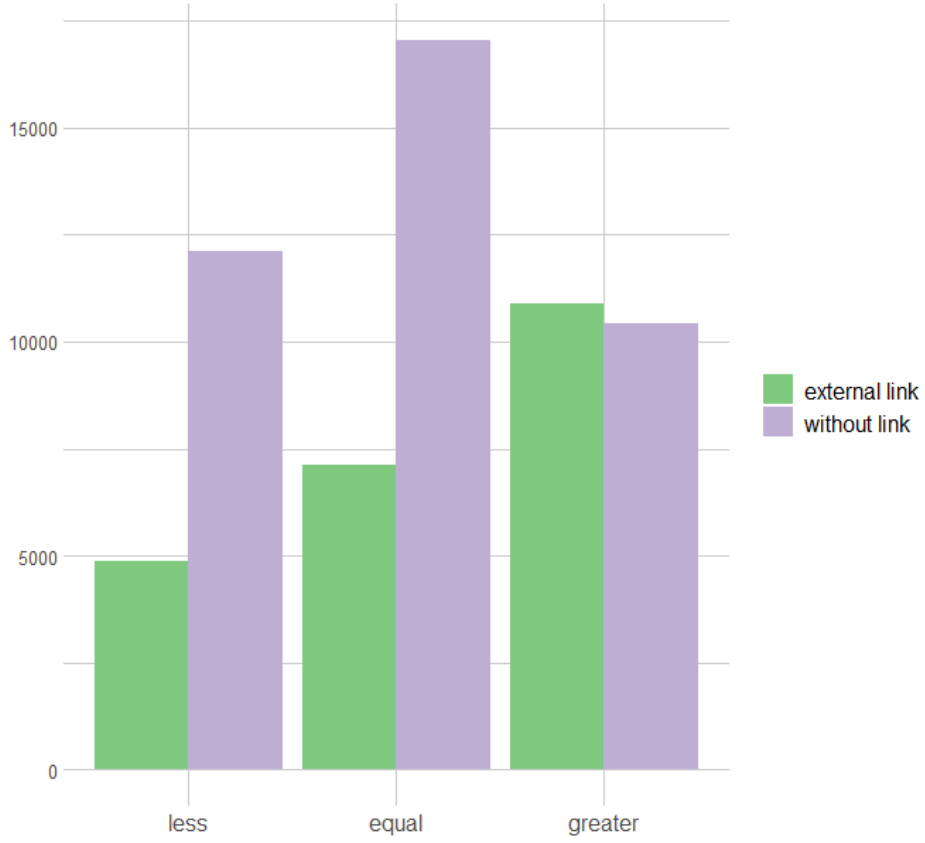

Figure 3. Publications in Semantic Scholar with higher, equal or fewer references than in Crossref according to whether or not they have an external link.

Semantic Scholar, however, offers a different perspective. With even more references than Crossref (3.34), it is the most comprehensive database. Semantic Scholar functions as a search engine by parsing references from publicly accessible online versions. This enables it to gather references that are not stored in Crossref, although it is also prone to mistakes in reference extraction and identification. Figure 3 shows the number of Semantic Scholar publications with fewer, equal and more references than Crossref, depending on whether or not the paper has a web link to an external copy. The figure indicates that publications with external links typically contain more references (51.1%), but the percentage of articles with external links that have the same or fewer references (28.5%) tends to be substantially lower. In other words, the likelihood of capturing more references than Crossref nearly doubles when an external copy is found. In Semantic Scholar, the quantity of references and external connections is associated,

as confirmed by the X-square test ($\chi 2=2890$, p-value<.001). This finding suggests that the references that Semantic Scholar extracts from the Web are being overstated. A manual inspection of 746 references from 15 publications with more references than the original papers showed that a large part of these extra references (59) are publications with similar titles or authors (33, 55.9%), are different versions of the same publication (pre-print, book instead of book chapter, conference, etc.) (23, 39%) or are duplicated references (3, 5.1%). Furthermore, we have observed that 6,966 (7.5%) papers have references that were published a year after the paper's publication date, indicating the inclusion of incorrect references.

5.4. References by document type

Document types could play a role in determining differences in the coverage of references, because each type of publication would include references in different forms, which could require more or less processing effort.

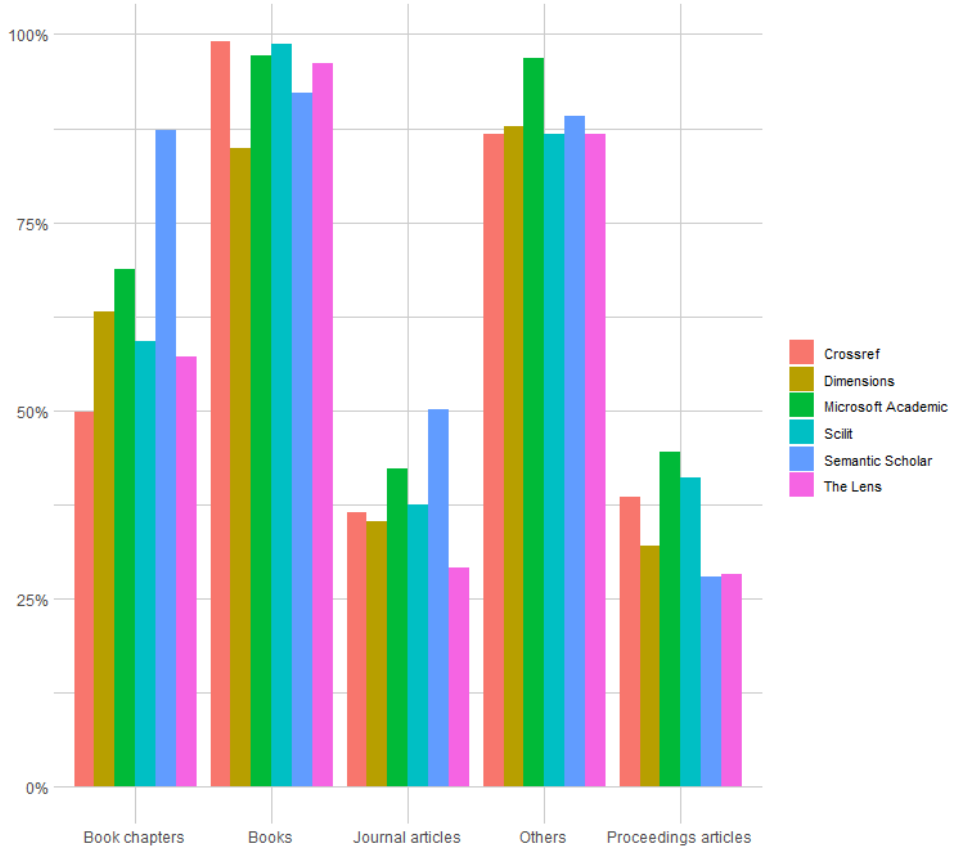

Figure 4. Proportion of document types without references in each database.

	Book chapters		Books		Journal articles		Others		Proceedings articles	
	No ref.	%	No ref.	%	No ref.	%	No ref.	%	No ref.	%
Crossref	7,192	49.8%	1,615	99.0%	31,857	36.5%	1,710	86.8%	3,852	38.5%
Dimensions	9,138	63.2%	1,392	84.9%	30,789	35.3%	1,726	87.7%	3,206	32.0%
Microsoft Academic	9,943	68.8%	1,594	97.1%	36,959	42.4%	1,905	96.8%	4,450	44.5%
Scilit	8,558	59.2%	1,619	98.7%	32,699	37.5%	1,706	86.7%	4,108	41.1%
Semantic Scholar	12,614	87.3%	1,512	92.1%	43,704	50.1%	1,754	89.2%	2,796	27.9%
The Lens	342	57.1%	75	96.2%	1,090	29.2%	79	86.8%	123	28.3%

Table 3. Proportion of publications by document type without references in each database.

Figure 4 and Table 3 display the percentage of publications, broken down by database, without references. These publications have been grouped by five primary document typologies from Crossref: Books, Book chapters, Journal articles, Others and Proceedings articles. The figure clearly shows that references from Books are less covered by each database, ranging from the 84.9% of Dimensions to the 99% of Crossref. In the case of Others, with 89% on average, it is explained because this class contains documents without references, such as peer-review, datasets or components. In contrast, the types of papers with the highest indexing of references are journal articles (38.5% on average) and proceedings articles (35.4% on average). The results also evidence that all the databases describe similar proportions. We can thus highlight the significant percentage of journal articles (50.1%) and book chapters (87.3%) in Semantic Scholar that lack citations. High scores are also reported by Microsoft Academic for Proceedings articles (44.5%), Others (96.8%) and Book chapters (68.8%).

5.5. References by publisher

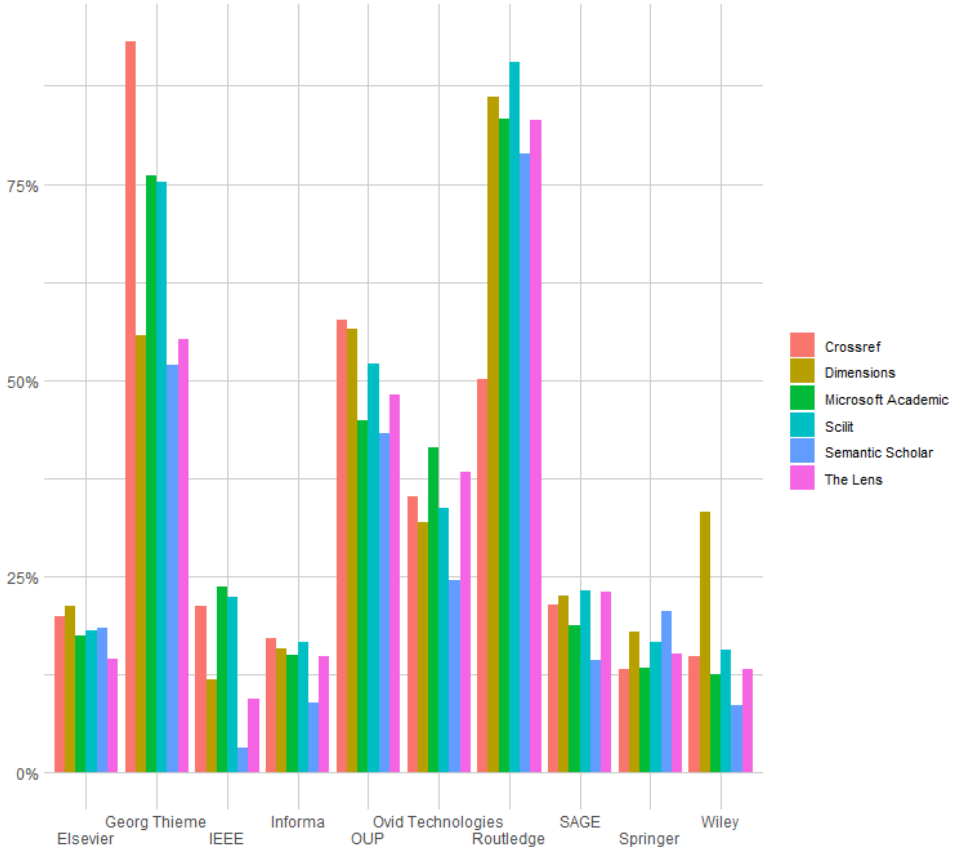

Figure 5. Proportion of publications without references in each database and breakdown by publisher.

Publishers	Crossref		Dimensions		Microsoft Academic		Scilit		Semantic Scholar		The Lens	
	No ref.	%	No ref.	%	No ref.	%	No ref.	%	No ref.	%	No ref.	%
Elsevier	3875	19.9%	4144	21.2%	2953	17.4%	3283	18.1%	1442	18.5%	116	14.5%
Springer	1841	13.2%	2502	17.9%	1561	13.4%	2302	16.6%	1021	20.6%	89	15.1%
IEEE	1533	21.3%	851	11.8%	1497	23.8%	1585	22.3%	194	3.2%	30	9.3%
OUP	1527	57.7%	1496	56.5%	903	44.9%	1259	52.1%	900	43.2%	53	48.2%

Wiley	898	14.8%	2017	33.3%	634	12.6%	941	15.6%	256	8.6%	36	13.1%
Routledge	898	50.1%	1545	86.2%	589	83.3%	1622	90.5%	851	78.9%	69	83.1%
Informa	676	17.2%	621	15.8%	534	15.0%	654	16.6%	333	8.9%	26	14.9%
SAGE	398	21.5%	419	22.6%	296	18.7%	430	23.2%	208	14.3%	17	23.0%
Georg Thieme	673	93.2%	402	55.7%	424	76.1%	542	75.2%	300	52.0%	16	55.2%
Ovid Technologies	573	35.2%	519	31.9%	550	41.4%	547	33.7%	314	24.5%	28	38.4%

Table 4. Proportion of publications by publisher without references in each database.

Figure 5 and Table 4 display the percentage of papers lacking references according to the ten publishers in the Crossref sample with most publications. As can be seen in the figure, the average percentage of publications without references from Routledge (78.7%), Georg Thieme (67.9%) and Oxford University Press (OUP) (50.5%) is higher than that of Informa (14.7%) and IEEE (15.3%). This result clearly shows that the way publishers distribute this information—both by depositing it in Crossref and making it accessible online—influences how references are indexed. Some of these publishers specialize in book publishing. For example, publications by Routledge in the sample consist entirely of books and book chapters. However, these types are only a small part of the Georg Thieme (19.5%) and OUP (27%) publications, which suggests that this lack of references is not solely explained by document typologies but also by the way in which publishers release these data.

6. Discussion

Comparative analysis of scholarly databases, with respect to the number and type of bibliometric indicators, revealed that only two databases—Dimensions and Semantic Scholar—develop advanced indicators that measure different facets of citation impact. Dimensions, in particular, uses altmetric indicators (Altmetric Attention Score) that establish links with societal impact and relative indicators (FCR, RCR) that contextualize the citation value. Additionally, Semantic Scholar provides structural indications (Influential Citation Count) that enhance the citation's meaning. All of these indicators align with the latest research evaluation trends (DORA, CoARA), which demand more elaborated metrics that capture more contextualized impact. However, the remaining databases only compute raw citations, perpetuating the accumulative notion of scientific impact. This leads us to the conclusion that, rather than being employed as instruments for research assessment, many of these new databases are better suited as discovery tools.

However, the validity of these indicators resides in how citations are processed. Significant differences in citation computation have been found through comparison analysis, primarily by academic search engines. Thus, Crossref-based databases (Crossref, Dimensions, Scilit and The Lens) compute comparable citation counts because they rely on the same citation source but also because they utilize the DOI as an unambiguous document identifier. Contrarily, Google Scholar introduces the notion of *version* to group different documents with the same content and, in consequence, to aggregate citations to the same cluster of records. This distorts the comparison and introduces biases in the evaluation of these documents because this bias favours publications with multiple editions such as books. It is not surprising that books make up 62% of Google Scholar's most cited documents (Martín-Martín et al., 2014). According to research on citation counts in Microsoft Academic, the results show that this product may suffer from a lack of update since 2017 (Visser et al., 2021; Scheidsteger & Haunschild, 2023), which indicates that updating is essential for accurate citation calculations.

Aside from these issues with updating and aggregating citations, another important consideration in citation computation is the accurate coverage of bibliographic references. The original aspect of this study is the introduction of references analysis as a means of elucidating variations in citation counts. We have thus seen that the primary cause of the negative fluctuation in citations in Scilit was mainly due to the low amount of indexed references from Crossref, caused in turn by the fact that Scilit only indexed DOI-based references. In the same manner, this reference analysis has also helped us to detect why Semantic Scholar is the second database with the most citations (behind Google Scholar). The reason is the elevated amount of indexed references in publications with external web copies, which suggests that Semantic Scholar could wrongly include references from different documents or could duplicate the references. These issues draw attention to the diverse challenges that third-party databases (Scilit) and search engines (Semantic Scholar) encounter when indexing references. Third-party databases only collect references that match the indexed publications, which would explain why Crossref, the original source, has more references than Dimensions, Scilit and The Lens. Those references cannot be incorporate to the system because they do not include pertinent information about authors, organizations and disciplines. On the contrary, some search engines such as Google Scholar do update their databases with new references, which make up 18% of the entire database (Orduña-Malea et al., 2015). This approach sacrifices metadata richness for coverage. However, the main challenge of academic search engines is parsing references. As we have seen, Semantic Scholar has significant problems when attempting to compile references from different web versions. A problem that has also been reported in Microsoft Academic (Visser et al., 2021) and Google Scholar (Martín-Martín et al., 2016).

Nevertheless, databases are also subject to external influences that affect the coverage of references and, consequently, the number of citations. The findings have demonstrated that not all databases have appropriately indexed book and book chapter references. This could have important consequences when citations are calculated for disciplines such as humanities and social sciences because they make considerable use of these materials. This problem is added to the incomplete coverage of books and book chapters (Delgado-Quirós et al., 2024) and low-quality metadata on these document types (Delgado-Quirós & Ortega, in press). A significant portion of these limitations are down to the publishers themselves because they neither properly add the references to Crossref nor facilitate their extraction from the Web. Our results show that references from publications by Georg Thieme, Routledge and, to a lesser extent, Oxford University Press, have trouble being indexed in all databases. This fact emphasizes the importance of scholarly publishers, and particularly book publishers, providing bibliographic references and ensuring that they are correctly indexed in scholarly databases.

Methodologically, the use of average variations instead of correlations has proven to be more reliable when differences between databases are analyzed. The advantage is that it makes it possible to show the direction of the differences, indicating which database computes the most or least number of citations on average. Moreover, this indicator also helps to improve the assessment of the extent of these differences. The main drawback of this measure is that it makes sense for skewed distributions, such as citation counts. Consequently, rather than being used as a precise parameter, these average variations should be sought as an approximate estimate.

6.1. Limitations

The comparative study of bibliographic databases has always had an intrinsic limitation when the benchmark database is selected because the comparison is always determined by the biases of the control database. In our case, Crossref is limited to publications deposited by specific consortium members and is not an exhaustive collection of scientific publications. Another limitation is that the deposited information is not always complete, and some publishers do not include accurate information about references, as we have seen.

Several data extraction processes (API REST, web scraping, dump files, etc.) were followed according to each database. This disparity of means has been able to cause data errors and inaccuracies, mainly in the use of web scraping and searches by title. However, the pairwise comparison of publications has palliated this limitation because we have only compared publications included in both databases.

Another possible limitation is related to the selection of sources and the changing environment in which academic databases are involved. Data were obtained at the same time in 2021, and since then, products such as Microsoft Academic have disappeared, while new databases such as OpenAlex have emerged. Thus, the picture obtained may not be entirely representative of the current panorama, because these products have been able to include references that were not previously indexed, and/or to improve the accuracy computing citations. Then, it is possible that some of the mistakes computing citations and processing references could have been mitigated. We welcome more comparative approaches to monitoring the reliability of these products regarding citation counts because the handling and processing of citations by scholarly databases is a critical element for their consideration for research evaluation.

7. Conclusions

The results of this investigation lead us to conclude that, with the exception of Dimensions and Semantic Scholar, the databases examined incorporate citation counts solely as an impact measure, which does not make it possible to value the performance of scholarly outputs in a precise context. This fact reveals that the primary purpose of these new scholarly databases is to serve as instruments for searching academic material rather than for research evaluation exercises. Therefore, we conclude that Scilit, The Lens or Google Scholar are not recommendable for use in research evaluation.

Comparative analysis of citation counts has revealed that there are significant variations according to the number of computed citations. These differences are less significant across Crossref-based tools, but noteworthy in academic search engines. These discrepancies are mostly caused by Microsoft Academic's upgrading issues and the integration of many copies (versions) in Google Scholar and Semantic Scholar. The study on reference coverage has also shown that Semantic Scholar has considerable difficulty parsing references from web copies, and Scilit has limits when it comes to incorporating references.

Finally, the study of references has also led us to detect how specific document types, such as books and book chapters, present limitations when their references are extracted and indexed in scholarly databases. The results suggest that these issues are caused by publishers, who fail to make the references available in a suitable way. In conclusion, scholarly publishers should exhibit greater transparency when releasing bibliographic references because it affects the findability of publications and their scientific impact.

8. Competing interests statement

José Luis Ortega is a member of the Scilit advisory board, which facilitated access to API documentation.

9. Author contributions

Lorena Delgado-Quirós has contributed to Data curation, Formal analysis, Resources and Software. José Luis Ortega has contributed to Writing, Conceptualization, Investigation, Methodology and Funding acquisition.

10. Funding information

This work was supported by the research project (NewSIS) “New scientific information sources: analysis and evaluation for a national scientific information system” (Ref. PID2019-106510GB-I00) funded by the Spanish State Research Agency (AEI) PN2019.

11. References

Archambault, É., Campbell, D., Gingras, Y., & Larivière, V. (2009). Comparing bibliometric statistics obtained from the Web of Science and Scopus. *Journal of the American society for information science and technology*, 60(7), 1320-1326.

Ball, R., & Tunger, D. (2006). Science indicators revisited–Science Citation Index versus SCOPUS: A bibliometric comparison of both citation databases. *Information Services & Use*, 26(4), 293-301.

Bauer, K., & Bakkalbasi, N. (2005), An examination of citation counts in a new scholarly communication environment. *D-Lib Magazine*, 11(9), http://www.dlib.org/dlib/september05/bauer/09bauer.html

Boyle, A. (2018). AI2 joins forces with Microsoft Research to upgrade search tools for scientific studies. *GeekWire*. https://www.geekwire.com/2018/ai2-joins-forces-microsoft-upgrade-search-tools-scientific-research/

Carpenter, M. P., & Narin, F. (1981). The adequacy of the Science Citation Index (SCI) as an indicator of international scientific activity. *Journal of the American Society for Information Science*, 32(6), 430-439.

Delgado-Quirós, L., Aguillo, I. F., Martín-Martín, A., López-Cózar, E. D., Orduña-Malea, E., & Ortega, J. L. (2024). Why are these publications missing? Uncovering the reasons behind the exclusion of documents in free-access scholarly databases. *Journal of the Association for Information Science and Technology*, 75(1), 43-58.

Delgado-Quirós, L., & Ortega, J. L. (2024). Completeness degree of publication metadata in eight free-access scholarly databases. *Quantitative Science Studies*, 5(1), 31-49.

Delgado-Quirós, L., & Ortega, J. L. (in press). Research entity information and coverage in eight free access scholarly databases. *Online Information Review*. https://osf.io/n7gsh

Gallagher, E. J., & Barnaby, D. P. (1998). Evidence of methodologic bias in the derivation of the Science Citation Index impact factor. *Annals of emergency medicine*, 31(1), 83-86.

Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, 122(3159), 108-111.

Guerrero-Bote, V. P., Chinchilla-Rodríguez, Z., Mendoza, A., & de Moya-Anegón, F. (2021). Comparative analysis of the bibliographic data sources dimensions and scopus: an approach at the country and institutional levels. *Frontiers in Research Metrics and Analytics*, 5, 593494.

Gusenbauer, M. (2019). Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. *Scientometrics*, 118(1), 177-214.

Hannousse, A. (2021). Searching relevant papers for software engineering secondary studies: Semantic Scholar coverage and identification role. *IET Software*, 15(1), 126-146.

Harzing, A. W. (2019). Two new kids on the block: How do Crossref and Dimensions compare with Google Scholar, Microsoft Academic, Scopus and the Web of Science?. *Scientometrics*, 120(1), 341-349.

Hendricks, G., Tkaczyk, D., Lin, J., & Feeney, P. (2020). Crossref: The sustainable source of community-owned scholarly metadata. Quantitative Science Studies, 1(1), 414-427.

Herrmannova, D., & Knoth, P. (2016). An analysis of the Microsoft academic graph. *D-lib Magazine*, 22(9/10), 1.

Hook, D. W., Porter, S. J., & Herzog, C. (2018). Dimensions: building context for search and evaluation. *Frontiers in Research Metrics and Analytics*, 3, 23.

Hug, S. E., Ochsner, M., & Brändle, M. P. (2017). Citation analysis with Microsoft Academic. *Scientometrics*, 111, 371-378.

Jacsó, P. (2008). Google scholar revisited. *Online information review*, 32(1), 102-114.

Kacperski, C., Bielig, M., Makorthyk, M., Sydorova, M., & Ulloa, R. (2023). Examining bias perpetuation in academic search engines: an algorithm audit of Google and Semantic Scholar. *arXiv preprint* arXiv:2311.09969.

Kousha, K., & Thelwall, M. (2008). Sources of Google Scholar citations outside the Science Citation Index: A comparison between four science disciplines. *Scientometrics*, 74(2), 273-294.

Line, M. B. (1979). The influence of the type of sources used on the results of citation analyses. *Journal of documentation*, 35(4), 265-284.

López-Illescas, C., de Moya-Anegón, F., & Moed, H. F. (2008). Coverage and citation impact of oncological journals in the Web of Science and Scopus. *Journal of informetrics*, 2(4), 304-316.

MacRoberts, M. H., & MacRoberts, B. R. (1989). Problems of citation analysis: A critical review. *Journal of the American Society for information Science*, 40(5), 342-349.

Martín-Martín, A., Orduña-Malea, E., Ayllón, J. M., & López-Cózar, E. D. (2014). Does Google Scholar contain all highly cited documents (1950-2013)?. *arXiv preprint* arXiv:1410.8464.

Martín-Martín, A., Orduña-Malea, E., Harzing, A. W., & López-Cózar, E. D. (2017). Can we use Google Scholar to identify highly-cited documents?. *Journal of informetrics*, 11(1), 152-163.

Martín-Martín, A., Orduña-Malea, E., Ayllón, J. M., & López-Cózar, 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.

Martín-Martín, A., Orduña-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., Thelwall, M., Orduña-Malea, E., & López-Cózar, E. D. (2021). Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations. *Scientometrics*, 126(1), 871-906.

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: a comparative analysis. *Scientometrics*, 106, 213-228.

Narin, F. (1976). Evaluative bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity (pp. 206-219). Cherry Hill, NJ: Computer Horizons.

Nederhof, A. J. (1985). Evaluating research output through life work citation counts. *Scientometrics*, 7, 23-28.

Orduña-Malea, E., Ayllón, J.M., Martín-Martín, A., & López-Cózar, E. D. (2015). Methods for estimating the size of Google Scholar. *Scientometrics* 104, 931–949.

Ortega, J. L. (2014). Academic search engines: A quantitative outlook. Oxford: Chandos Publishing (Elsevier). ISBN: 9781843347910

Scheidsteger, T., & Haunschild, R. (2023). Which of the metadata with relevance for bibliometrics are the same and which are different when switching from Microsoft Academic Graph to OpenAlex? *Profesional de la información*, 32(2).

Vera-Baceta, M. A., Thelwall, M., & Kousha, K. (2019). Web of Science and Scopus language coverage. *Scientometrics*, 121(3), 1803-1813.

Visser, M., van Eck, N. J., & Waltman, L. (2021). Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *Quantitative Science Studies*, 2(1), 20-41.

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.