Check for updates

Received: 18 November 2022 | Revised: 2 October 2023 | Accepted: 10 October 2023

DOI: 10.1002/asi.24839

RESEARCH ARTICLE

JASIST WILEY

Why are these publications missing? Uncovering the reasons behind the exclusion of documents in free-access scholarly databases

Lorena Delgado-Quirós[1] | **Isidro F. Aguillo**[2] | **Alberto Martín-Martín**[3] | **Emilio Delgado López-Cózar**[3] | **Enrique Orduña-Malea**[4] | **José Luis Ortega**[1]

[1]Institute for Advanced Social Studies (IESA), CSIC, Knowledge Transfer and Innovation, UCO, Associated Unit to CSIC by IESA, Córdoba, Spain

[2]Cybermetrics Lab (IPP-CSIC), Madrid, Spain

[3]Facultad de Comunicación y Documentación, Colegio Máximo de Cartuja, Universidad de Granada, Granada, Spain

[4]Department of Audiovisual Communication, Documentation and History of Art, Universitat Politècnica de València, Valencia, Spain

Correspondence
José Luis Ortega, Institute for Advanced Social Studies (IESA), CSIC, Knowledge Transfer and Innovation, UCO, Associated Unit to CSIC by IESA, Camposanto de los mártires, 7, Córdoba 14004, Spain.
Email: jortega@iesa.csic.es

Funding information
Spanish State Research Agency (AEI), Grant/Award Number: PID2019-106510GB-I00

Abstract

This study analyses the coverage of seven free-access bibliographic databases (Crossref, Dimensions—non-subscription version, Google Scholar, Lens, Microsoft Academic, Scilit, and Semantic Scholar) to identify the potential reasons that might cause the exclusion of scholarly documents and how they could influence coverage. To do this, 116 k randomly selected bibliographic records from Crossref were used as a baseline. API endpoints and web scraping were used to query each database. The results show that coverage differences are mainly caused by the way each service builds their databases. While classic bibliographic databases ingest almost the exact same content from Crossref (Lens and Scilit miss 0.1% and 0.2% of the records, respectively), academic search engines present lower coverage (Google Scholar does not find: 9.8%, Semantic Scholar: 10%, and Microsoft Academic: 12%). Coverage differences are mainly attributed to external factors, such as web accessibility and robot exclusion policies (39.2%–46%), and internal requirements that exclude secondary content (6.5%–11.6%). In the case of Dimensions, the only classic bibliographic database with the lowest coverage (7.6%), internal selection criteria such as the indexation of full books instead of book chapters (65%) and the exclusion of secondary content (15%) are the main motives of missing publications.

1 | INTRODUCTION

Scholarly bibliographic databases are key elements to support the advance of science because they provide updated information about past scientific developments that makes possible to contrast current discoveries. Many of these products index the cited references included in the publications to enlarge the discovery of information and to value the influence of that records. Web of Science and Scopus are traditional citation indexes that gather bibliographic records from a selected list of sources, mainly scholarly journals.

However, the advent of the Web in 1989 meant the transformation of the publishing model (Borgman & Furner, 2002), and consequently, a new way to gather publications and measure citations. Launched in 1997,

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2023 The Authors. *Journal of the Association for Information Science and Technology* published by Wiley Periodicals LLC on behalf of Association for Information Science and Technology.

CiteSeer was the first academic search engine that used a crawler to harvest electronic publications, extracting and computing citations between publications (Fiala, 2011). This model served as basis for subsequent developments such as Google Scholar (Delgado López-Cózar et al., 2019) and Microsoft Academic (Wang et al., 2019, 2020). Search engine-based bibliographic information systems tend to provide more comprehensive document coverage than traditional selective systems, due to the digital transformation of the publishing system and the proliferation of repositories and web platforms (Ortega, 2014).

Currently, new hybrid models combining publications gathered both through traditional curation and web crawling have been released, thanks to improvements in the harvesting, storage, and processing of bibliographic data, besides the free releasing of citation metadata (Open Citations) (Ortega, 2021). The free version of Dimensions (Herzog et al., 2020; Orduña-Malea & Delgado-López-Cózar, 2018; Thelwall, 2018), Lens (Penfold, 2020), or Scilit are new hybrid information services (referred to as free-access databases) that are facilitating discovery of the scientific literature as well as providing new analytic tools and bibliometric indicators (i.e., altmetrics, field-normalized metrics, usage-based metrics).

The proliferation of new free-access scholarly databases has fostered many studies comparing their coverage and overlap (see section 2) to help practitioners, meta-researchers and scholars to select the most appropriate databases to carry out systematic literature reviews, meta-analysis, bibliometric analyses, or literature searches (Bramer et al., 2017; Gusenbauer & Haddaway, 2020). As Mongeon and Paul-Hus (2016) states, "the validity of bibliometric analyses for research evaluation lies in large part on the databases' representativeness of the scientific activity studied." An incorrect selection of scholarly databases might report incomplete or misleading results and false conclusions.

However, most of these analyses are based on the direct comparison of one database with the other ones. In our opinion, this procedure could sweep along biases from the original database and distorting the coverage of those sources. This study attempts to develop a different approach using a random sample from a non-selective service (Crossref) to compare different scholarly databases' coverage. We hypothesize that using a third party service, that is, using a third database to compare the coverage of other two, would reduce possible biases in the comparison of databases, as well as to know how selection criteria and technical requirements influence the coverage of scientific literature.

2 | LITERATURE REVIEW

As information products, scholarly databases can be evaluated under different features (e.g., search and results interface, quality of data, exporting capabilities), where coverage is one of the most important parameters to test their bibliometric capabilities. Coverage is measured not only to test the databases' power to find and index scientific literature, but also to check their completeness and to detect potential biases. Coverage can be measured in two different ways: measuring indexed documents and measuring cited documents (coverage via citations). Comparative analyses can be carried out by applying different methods (e.g., direct comparisons, third-party comparisons). The most relevant literature on free-access bibliographic databases coverage is discussed below.

2.1 | Coverage via direct comparisons

A way to test coverage biases is a direct comparison between databases, with the aim to identify the most appropriate database according disciplines, documents types, or languages.

Due to the scarce availability of data, the first studies on the topic were focused on cited documents (citations) as proxy of coverage (Bakkalbasi et al., 2006; Levine-Clark & Gil, 2008; Meho & Yang, 2007). All of them concluded that Google Scholar surpassed Web of Science and Scopus. Kousha et al. (2011) demonstrated that Google Scholar captured more citations to books and book chapters than traditional citation indexes, while Adriaanse and Rensleigh (2013) warned that the higher citation count of Google Scholar could be due to duplicated records, while other errors might occur due to the uncontrolled nature of the database (Orduña-Malea et al., 2017). The appearance of Microsoft Academic produced several studies that show this new product to perform similarly to Google Scholar (Haley, 2014; Harzing & Alakangas, 2017; Ortega & Aguillo, 2014), also improving the citation coverage of Web of Science and Scopus (Hug & Brändle, 2017).

Another approach is measuring indexed documents, using standardized search queries to compare the results in several platforms. Jacso (2005) was the first one to use several specific search queries to compare the coverage of several databases, finding that Google Scholar surpassed the coverage of Web of Science and Scopus. Khabsa and Giles (2014) used this procedure to estimate the size of Google Scholar (100 M), Microsoft Academic (50 M), Web of Science (50 M), and Pubmed (20 M). Orduña-Malea et al. (2015) employed the same method to estimate the size of Google Scholar, concluding an estimation of 160–165 million documents, a figure

23301643, 2024, 1, Downloaded from https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.24839 by Readcube (Labtiva Inc.), Wiley Online Library on [15/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

subsequently updated to 331 million documents, including publications, cited references and patents (Delgado López-Cózar et al., 2019). Later, Gusenbauer (2019) performed the largest comparison of scholarly databases counting query hits, calculating 389 million documents indexed in Google Scholar. The drawback of this method is that the results are always estimations and they depend on the search interface of each database.

The availability of data (e.g., API endpoints, dump files) and the proliferation of new products increased the number of coverage studies using direct comparisons. Van Eck et al. (2018) were the first one in comparing Crossref with traditional citation indexes (i.e., Web of Science and Scopus). Their results showed that Crossref had a similar coverage, but with limitations according to reference and metadata quality. Harzing (2019) concluded that Crossref and Dimensions could be good alternative to traditional citations indexes, but not for academic search engines such as Google Scholar and Microsoft Academic. Singh et al. (2021) adopted a journal coverage approach to compare Web of Science, Scopus and Dimensions. Their results showed that Dimensions is more inclusive in the journal indexation than the other platforms. Guerrero-Bote et al. (2021) compared Scopus and Dimensions at country and organizational level, finding that Dimensions lacked affiliation data in more than half of the publications. Finally, Purnell (2022) showed that large databases such as Dimensions and Microsoft Academic have more affiliation discrepancies than Scopus or Web of Science.

2.2 | Coverage via third party comparisons

The use of third party sources to compare the coverage of bibliographic databases is scarce. We can highlight the use of Google Scholar's *Classic Papers* product[1] as a baseline to generate comparisons between free access and traditional databases, measuring both indexed documents (Martín-Martín, Orduna-Malea, & Delgado López-Cózar, 2018) and citations (Martín-Martín et al., 2021; Martín-Martín, Orduna-Malea, Thelwall, & Delgado López-Cózar, 2018). Specifically, Martín-Martín, Orduna-Malea, and Delgado López-Cózar (2018) showed that a large fraction of highly-cited documents in the Social Sciences and Humanities (8.6%–28.2%) were invisible to Web of Science and Scopus. Martín-Martín, Orduna-Malea, Thelwall, and Delgado López-Cózar (2018) compared 2 M Google Scholar citations with Scopus and Web of Science according to disciplines, evidencing that Google Scholar detected 37% more citations than the traditional citation indexes. Martín-Martín et al. (2021) compared Google Scholar citations with other five bibliographic products, confirming that Google Scholar is the most comprehensive service finding citations.

However, these studies could be influenced by Google Scholar coverage, as the classic papers used to compare all the databases (2515 highly cited documents written in English and published in 2006) constituted a subset of Google Scholar, being all indexed in this database.

2.3 | Reasons for no indexation

Beyond the differences between scholarly databases, either measuring citing or cited documents, or using direct or third-party comparisons, very few studies had explored the reasons why these coverage differences occur. While testing her own curricula, Harzing (2019) observed which publications were not indexed in several platforms. Visser et al. (2021) manually checked the content of non-indexed documents in several sources envisaging that some of these documents did not contain scientific content. However, there are no studies whose objectives were focused on the causes of the no indexation.

3 | OBJECTIVES

The main objective of this article is to compare the coverage of the largest number of freely accessible databases (Dimensions, Google Scholar, Lens, Microsoft Academic, Scilit, and Semantic Scholar) using a third-party comparison (via Crossref) to show which databases differ in the coverage of publications, which allows us to identify potential reasons in the no indexation of data. Specifically, this study aims to answer the following research questions:

RQ1. Are there significant coverage differences among the currently available free-access bibliographic sources?

RQ2. Which document typologies cause greater coverage differences?

RQ3. What are the potential reasons behind the no indexation of documents in free-access bibliographic sources?

4 | METHOD

This study aims to select the most comprehensive range of scholarly databases, with the only limitation to sources that provide free access to their results. This means that

[1] https://scholar.google.com/citations?view_op=list_classic_articles&hl=en&by=2006.

23301643, 2024, 1, Downloaded from https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.24839 by Readcube (Labtiva Inc.), Wiley Online Library on [15/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

all the analyzed databases provide a search interface that makes possible to search and retrieve records without any cost (which excludes paywall citation indexes such as Web of Science and Scopus) and compute bibliometric indicators. For example, Dimensions and Lens could be considered *freemium* products that provide a free access to the search interface of a public version, but require a subscription or agreement to access to a version with more functionalities (i.e., Dimensions Plus, Lens Reports). In total, six bibliographic databases (Dimensions, Google Scholar, Lens, Microsoft Academic, Scilit, and Semantic Scholar) were analyzed according to a reference sample from Crossref.

4.1 | Crossref sample

This study takes a third-party approach, in which the comparison between databases is done through a third or control database. The strength of this procedure is its ability to avoid potential coverage biases in one database that could influence the comparison. Using a third-party database would reduce this risk because all the databases being compared are now influenced in the same way by the same external database, thus balancing the comparison.

Crossref was used as the control sample due to several reasons. The first reason is operational. This database is the main data provider of Document Object Identifiers (DOIs) for research publications, being the most extended persistent identifier of research publications in the publishing system.[2] Despite their coverage not being exhaustive (Visser et al., 2021), its use is justified because all the remaining six databases under study integrate the DOI as a searchable field, facilitating a rapid and exact matching. The second reason is methodological. Crossref allows the extraction of random samples of documents from its API and dump files. This fact favors the representativeness of the sample, avoiding ranking algorithms, filters, or matching procedures that could disrupt the coverage analysis. Random samples also reduce time and processing costs, favoring the comparison of multiple sources. The third reason is procedural. Crossref assigns DOIs to any published material in a book, journal, or conference, regardless of its informative value (e.g., front covers, indexes, news). Therefore, no inclusion criteria limit the coverage of certain types of documents. This non-selective criterion would lead us to clearly appreciate the inclusion policies of the other bibliographic databases.

[2]There are 11 registration agencies for DOIs (https://www.doi.org/RA_Coverage.html). Crossref includes more than 134 million of research publications in 2022, followed by DataCite, with 38 million of non-published materials. The rest of registration agencies are locals (i.e., Airiti, JaLC, KISTI) or cover non-scientific results (mEDRA, OP, EIDR).

4.2 | Data collection

A random sample of 116,648 DOIs from Crossref was retrieved in August 2020, and subsequently updated in July 2021. This sample was generated performing 1200 automatic requests to https://api.crossref.org/works?sample=100. This random process produced duplicate records that were removed to obtain the final list. These requests were limited to documents published between 2014 and 2018. The distribution by document type matches with the entire database (Hendricks et al., 2020), confirming the reliability of the sample. In addition, Table 1 compares some parameters of the entire Crossref database in May 2020 and the random sample in August 2020. The proportion of types of records is similar and constant around 0.1%, which confirms that the sample is balanced regarding the total database.

This control sample was subsequently queried to each database to match the records and extract all the information related to each publication. This task was performed during July 2021. A detailed description of the extraction process for each database and additional information (size and sources) is offered in Table 2.

In addition to conducting a search using DOIs, we have also conducted a search using the title of the publication on both Microsoft Academic and Google Scholar. The reason for searching on Microsoft Academic using the title is due to the low indexation rate of DOIs (37.1%). Consequently, we resorted to downloading the complete table of publications from Zenodo (https://zenodo.org/record/2628216) and matching the publications by their titles. As for Google Scholar, a search by title was carried out as there is no specific search option for DOIs. This was done to verify if additional publications could be retrieved by conducting a title search. The results showed that only 898 (0.8%) publications were identified. This suggests that the benefits of conducting a title search are minimal in comparison to the required efforts. Other databases, however, were not tested since their endpoints do not provide full title search or because they use Crossref as their main source.

Additional data processing was performed to explain the coverage of specific document types. For instance, to check the coverage of the entire book instead the book chapter, we had to remove the chapter suffix (e.g., https://doi.org/10.1002/9781119160243.ch3) or to search for the title of the book in the Web, and then extract its DOI. The categorization of secondary content

23301643, 2024, 1, Downloaded from https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.24839 by Readcube (Labtiva Inc.), Wiley Online Library on [15/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 Comparison between the total coverage of Crossref in May 2020 and the random sample (July 2021)

	Crossref	Sample	%
Total number of records in the database	125,094,590	116,648	0.09%
Total number of prefixes	20,343	5753	28.28%
Number of Journal DOIs	89,115,913	87,115	0.10%
Number of Book DOIs	19,751,190	16,428	0.08%
Number of conference DOIs	6,846,838	10,145	0.15%
Number of components	4,932,308	270	0.01%
Number of articles with references Deposited	51,260,467	34,105	0.07%

was done from the title of the document and exploring its content in their landing page.

4.3 | Identification of non-indexation reasons

We have defined the following two main types of reasons:

- Internal requirements: Each database defines what materials shall be indexed. These criteria could be motivated by informative (some documents could be more interesting to scholarly audiences), technical (some document types could require additional fields), or accessing reasons (some publications could not be openly available). For example, Google Scholar only indexes "scholarly articles," excluding "news or magazine articles, book reviews, and editorials" (Google Scholar, 2022), and Dimensions includes articles "from a scientific journal or trade magazine, including news and editorial content" (Dimensions Plus, 2019). Internal requirements are more associated to document types.
- External criteria: These conditions are caused by external sources that do not provide the information as the database requires. That is, the database decides including information but the source does not provide sufficient information (e.g., metadata) to be indexed. This problem is especially important in academic search engines, which use bots to crawl the Web and they require that the information is suitable for data harvesting. For example, we find the Google Scholar's *Inclusion Guidelines for Webmaster* (Google Scholar, 2022). External criteria are related to specific sources such as data providers or publishers.

Due to the particular operating mode of these external criteria, we adopted a web crawler perspective to identify the indexation problems. To do this, documents not found on Google Scholar, Microsoft Academic and Semantic Scholar were resolved using their DOI (12,404) (https://hdl.handle.net/) to explore the landing page of each publication. Next, a link checker (*Xenu's Link Sleuth*[3]) was used to test the accessibility of these webpages. Only those pages that returned the 200 status code (OK) were selected to be crawled, while the remaining ones were classified as access problems. An R script was written to extract robots' instructions (i.e., meta name = "robots", {{ngMeta.robots}}) and directions for robots' exclusion (i.e., noarchive, noindex).

This distinction of criteria allows to identify the principal reasons for not indexing specific documents, and to analyze the coverage problems in bibliographic databases and academic search engines in a differentiate way.

Data are available in the following URL: https://osf.io/h7yz9/.

5 | RESULTS

5.1 | Coverage

To compare the performance of each database and highlight the differences in the coverage of Crossref publications, the number and percentage of missing documents in each database is shown in Table 3. Lens (0.1%) and Scilit (0.2%) almost exactly reproduce the initial sample. On the other hand, Microsoft Academic (12%) and Semantic Scholar (10%) are the databases that miss more publications from our sample. The high missing values achieved by Dimensions (7.6%) and Google Scholar (9.8%) were unexpected. First, Crossref feeds Dimensions; and second, Google Scholar is considered the largest academic database (Gusenbauer, 2019; Martín-Martín et al., 2021). The high percentage of duplicated DOIs in Microsoft Academic is also worthy of mention (1107, 1.1%), and might be caused by the assignation of the same DOI to preprint copies and book chapters. Overall, these results show a high degree of overlap (>85%) with regard to the initial sample, confirming a high overlap between scholarly databases, a

[3]http://home.snafu.de/tilman/xenulink.html.

23301643, 2024, 1, Downloaded from https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.24839 by Readcube (Labtiva Inc.), Wiley Online Library on [15/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 2 Data collection process carried out in each bibliographic database

Database	Data collection	Size (millions)	Sources
Dimensions	Dimensions Analytics was used to collect the records through the API (https://docs.dimensions.ai/dsl). This restricted access was used only to facilitate the download of data, but no content not free-openly offered was analyzed. One of the co-authors has granted access to Dimensions Analytics. A R package (i.e., dimensionsR (https://github.com/massimoaria/dimensionsR) was used to extract the data. JSON format was used to download the results because dimensionsR caused some problems in the transformation of JSON outputs to CSV format	130	PubMed, PubMed central, Crossref + directly from content publisher
Google Scholar	As GS does not facilitate access to its data, web scraping was used to automatically query each DOI in the search box. The RSelenium R package (https://docs.ropensci.org/RSelenium/) was used to emulate a browser session and avoid anti-robot actions (i.e., captchas). As it is possible that some DOIs could be not indexed (Martín-Martín, Orduna-Malea, & Delgado López-Cózar, 2018), a title search with the query "allintitle:title" was used to complete the results	400	Directly from content publisher
Lens	After a formal request, this service provided us temporary access to its API (https://api.lens.org/scholarly/search). In this case, a R script was written to directly extract the data	247	Microsoft Academic, Crossref, Pubmed, Core, Open Alex
Microsoft Academic	First, SPARQL (https://makg.org/sparql) and REST API (https://api.labs.cognitive.microsoft.com/academic/v1.0/evaluate) endpoints were used to extract publications using DOIs. Then the entire table of publications available in Zenodo) was downloaded and locally matched with the sample, using now DOIs and titles	203	Directly from content publisher
Scilit	This platform was accessed using an internal API (https://www.scilit.net/api/v1) under request. An ad hoc R script was written to extract the data	149	Crossref + Pubmed
Semantic Scholar	This database provides a public API (https://api.semanticscholar.org/v1). The semscholar R package (https://github.com/njahn82/semscholar) was used to extract the data. However, API was directly queried after to detect some problems in the retrieval process	205	Directly from content publisher

Note: Estimated values as of August 26, 2022.

fact already found in the literature through other databases (Harzing, 2019; Visser et al., 2021).

The coverage differences shown in Table 3 are subsequently analyzed in the following sections to identify the reasons behind the no indexation of documents in certain databases. This tour allows us to uncover how methodological differences in the building, design and data feeding of these databases influence the indexation of scholarly publications.

5.2 | Similarities and differences

A first step to understand the different coverages is to study the similarities and differences among databases according to the overlap of documents. This overlap was calculated comparing the same records retrieved from the Crossref sample in each database. Figure 1 shows a multidimensional scaling (MDS) plot, in which the distances between services are calculated according to the proportion of overlapped documents in each platform. MDS was proposed to overcome the limitation of Venn diagrams of plotting more than three sets. K-means clustering algorithm was used to confirm the clusters (node color) observed in the MDS map. MDS is a visualization technique for displaying the information contained in a distance matrix. *K*-means is a clustering algorithm that groups elements according to the nearest mean of each cluster.

The *K*-means algorithm identifies an initial group (blue) shaped by Scilit and Lens, very similar to Crossref. This closeness evidences that both databases feed on Crossref incorporating almost entirely all the records stored in Crossref (miss <1%). The main characteristic of

23301643, 2024, 1, Downloaded from https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.24839 by Readcube (Labtiva Inc.), Wiley Online Library on [15/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 3 Missing publications of the Crossref random sample ($N = 116{,}647$) over the different scholarly databases

Data source	Duplicates	Duplicates %	Missing publications	Missing publications %
Crossref	0	0.00%	0	0%
Dimensions	35	0.03%	8860	7.6%
Google Scholar	0	0.00%	11,376	9.8%
Lens	0	0.00%	24	0.1%
Microsoft Academic	1107	1.08%	13,985	12%
Scilit	123	0.11%	259	0.2%
Semantic Scholar	0	0.00%	11,872	10%

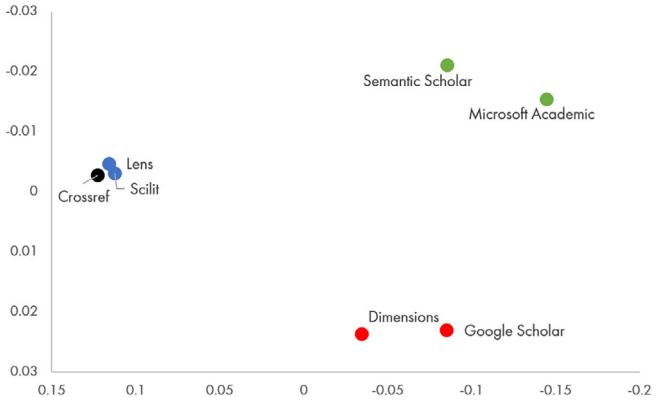

FIGURE 1 MDS map showing the distance between scholarly databases regarding to the overlap of indexed documents

this group is that these products create their databases incorporating publications from external data sources (i.e., Crossref, PubMed, Microsoft Academic).

A second intermediary group (red) is set by Dimensions and Google Scholar (miss ≈10%). Both databases have different methodologies to create their databases. While Dimensions is also based on external sources, Google Scholar is mainly supported on web crawling. Their similarity could be due to their broad coverage.

A third group (green), very far from Crossref (>10%), is shaped by Microsoft Academic and Semantic Scholar, two similar exact academic search engines that obtain their information through web crawlers. This similarity is also explained because Semantic Scholar also uses Microsoft Academic as data source (i.e., Microsoft Academic Graph) (Boyle, 2018).

These results evidence that the methodological and technical approaches used in the building of scholarly databases influence to great extent the coverage of documents.

5.3 | Missing publications

Figure 2 shows the proportion of missing publications by document type in comparison with Crossref. The aim is to check whether the document typology has any influence in the coverage, and to highlight which specific types of documents are prone to be indexed in each database. Due to this, Crossref is included in the graph to contrast the proportion of indexed documents in this database with the proportion of missing documents in the other ones. Crossref categories were used as reference in the comparison. These were grouped in eight principal classes: book (includes monographs and book series), book chapter (includes reference entry, reference book), dataset, journal article (includes journal issue), posted content, proceedings (include proceedings article), report, and other (includes component, correction, retraction, peer review). Appendix A details the number and proportion of all document typologies (Table A1). Overall, the bar graph shows that some typologies, in particular book chapters and journal articles, experience more problems to be indexed (Figure 2).

Figure 2 also shows different patterns according to the type of database. This way, in databases mainly based on Crossref (e.g., Scilit and Lens), most of the documents from the sample that are not found are journal articles (Scilit: 76.1%, Lens: 83.3%). However, these percentages are similar to the total coverage of journal articles in Crossref (74.7%), which suggest that this lack of coverage could not be due to this specific document typology. Academic search engines (Google Scholar, Microsoft Academic, and Semantic Scholar) show a different pattern, finding difficulties in the indexation of journal articles and book chapters in a similar proportion. For instance, of the documents in the sample that Google Scholar does not index, 41.3% are book chapters and 43.5% are journal articles. A manual inspection of a random sample of these documents ($N = 1354$; Confidence interval = 95%; error margin = 2.5%) disclosed that only 63.7% of the book chapters had scholarly content and 17.3% were reference entries. In the case of journal articles, where there are more different types, only 32.8% were strictly research papers. In Microsoft Academic, book chapters account for 39.9% of the documents that are not found, and journal articles make up 42.9% of the

23301643, 2024, 1, Downloaded from https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.24839 by Readcube (Labtiva Inc.), Wiley Online Library on [15/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

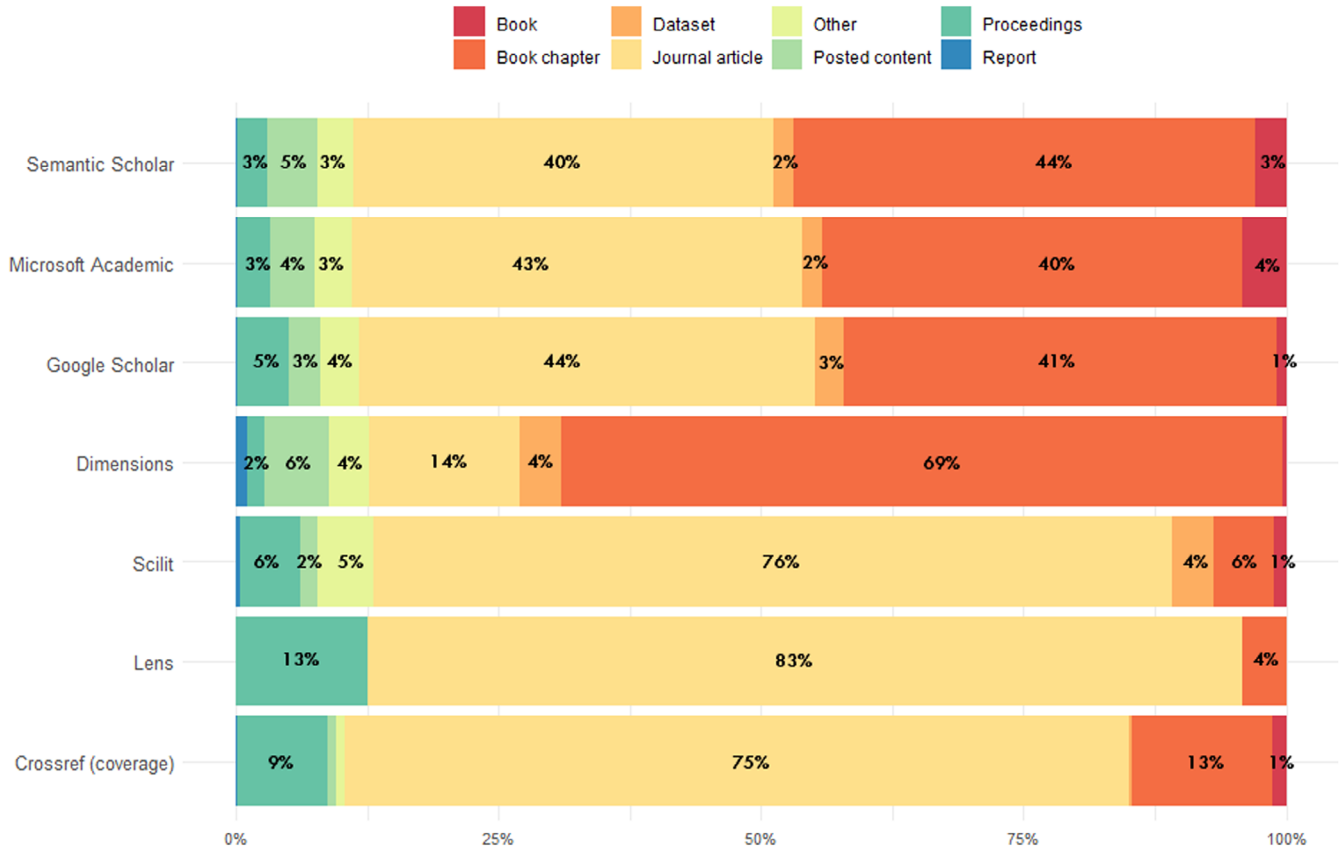

FIGURE 2 Distribution of non-indexed documents by typology in each platform

missing documents; and in Semantic Scholar, 43.91% of the missing documents are book chapters and 40% journal articles. Finally, Dimensions displays a particular pattern, finding problems specifically in the indexation of book chapters (68.8% of missing documents).

5.4 | Reasons for non-indexation

Next, we analyze why certain document types experience more problems to be indexed and how the different databases manage to index them.

5.4.1 | Bibliographic databases

The coverage of these databases is mainly determined by internal requirements. In the case of Lens (99.9%) and Scilit (99.8%), we consider that there are no inclusion criteria regarding Crossref data because the coverage is almost complete. In the case of Scilit, it is worth mentioning that 61.4% of the missing publications are records without a title, which suggests that both Lens and Scilit only employ technical criteria to exclude content, such as metadata completeness of the records.

TABLE 4 Principal causes for missing records from Crossref in Dimensions

Causes	Publications	Publications %
Full books	5796	65.4%
Secondary content	1357	15.3%
Posted content	506	5.7%
Datasets	348	3.9%
Oxford University Press	200	2.3%
Reports	91	1.0%
Other	523	5.9%
Total	8860	100%

Beyond internal indexing criteria, we find additional causes to explain the non-indexation of documents. Considering Dimensions as a case study (Table 4), which also uses Crossref as primary source, we find the following causes:

- Book chapters (65.4%): Book chapters are not separated from the full book. In other words, despite the full book being indexed, some of their chapters are missing. This problem occurs with 37.3% of the book chapters in the whole sample.

23301643, 2024, 1, Downloaded from https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.24839 by Readcube (Labtiva Inc.), Wiley Online Library on [15/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

- Secondary content (15.3%): Secondary content refers to publications that have a DOI but, strictly, they are not research publications. For example, in the case of journal articles, we can find editorials, news, table of contents, front matters, covers, etc., that accompany research articles but they do not have scientific content in their own. In the case of book chapters, this secondary content is related to indexes, forewords, abbreviations, glossaries, etc. 30.8% of all the secondary content in Crossref is excluded from Dimensions, while the remaining ones correspond to news and editorials that are indeed indexed.
- Posted content (5.7%): Dimensions eliminates post publication comments from Copernicus (88.2%) and abstracts from Morressier (100%).
- Datasets (3.9%) and Reports (1%): Are other formats that are almost entirely excluded.

It is worth mentioning that more than half of books and book chapters from *Oxford University Press* (50.1%) are not indexed, which suggests that Dimensions experiences certain problems when it comes to indexing bibliographic data from this publisher.

5.4.2 | Academic search engines

This group refers to scholarly information databases that mainly use crawlers and bots to gather bibliographic information. 10,554 webpages (85.1%) returned 200 (Ok) status code, being the remaining ones classified as "access" problems. Otherwise, only 5283 (50.1%) webpages had instructions for robots and 4819 (91.2%) included directions for robot exclusion.

Table 5 depicts the main causes that explain why some Crossref publications are not indexed in the academic search engines under analysis. Notice that some of these criteria are different from scholarly databases, illustrating the important methodological differences in the construction of these products. The distribution of causes found in the three search engines show similar percentages, suggesting that some of these external criteria equally influence each search engine. However, it is important to notice that these causes only explain 88.3% of missing documents in Semantic Scholar, 87.8% in Google Scholar, and 87.9% in Microsoft Academic.

As with Dimensions, the main limitation to index publications in Google Scholar is the indexation of entire books instead of the chapters (38.3%), an issue that affects 28% of the book chapters in the sample. Although it is officially stated that "Google Scholar automatically includes scholarly works from Google Book Search" (Google Scholar, 2022), the chapters of these books are not disaggregated, and then these documents cannot be retrieved from Google Scholar, unless the author/publisher has uploaded the specific chapter to some source indexed by Google Scholar. This problem is also remarkable in Microsoft Academic (19.3%) and Semantic Scholar (31.3%), although the reasons are unknown, and they could be due to the inexistence of appropriate landing pages or insufficient information for indexing them correctly.

The second most important case for no indexation is the robot exclusion. This is the main external limitation that prevents indexing of publications in academic search engines. Google Scholar (24%) is the service less affected by this problem, while this is the principal reason in Microsoft Academic (35.2%) and Semantic Scholar (35.2%). The better performance of Google Scholar in this area may have to do with how it indexes documents that it finds in the lists of cited references of other documents

TABLE 5 Principal causes for the missing records from Crossref in academic search engines

Academic search engines	Google Scholar		Microsoft Academic		Semantic Scholar	
Causes	Publications	Publications %	Publications	Publications %	Publications	Publications %
Full books	4355	38.3%	2703	19.3%	3648	31.3%
Robot exclusion	2728	24.0%	4918	35.2%	4106	35.2%
Access problems	1060	9.3%	1114	8.0%	777	6.7%
Secondary content	745	6.5%	1629	11.6%	1112	9.5%
No abstract	670	5.9%	401	2.9%	0	0.0%
Dataset, posted content	433	3.8%	676	4.8%	647	5.6%
Others	1385	12.2%	2544	18.2%	1366	11.7%
Total	11,376	100%	13,985	100%	11,656	100%

23301643, 2024, 1, Downloaded from https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.24839 by Readcube (Labtiva Inc.), Wiley Online Library on [15/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

(i.e., Citations), and how it is able to identify different versions of the same document in different websites, both of which are practices that could lessen the impact of a robot exclusion policy in a particular website. Another external problem is the open availability of publications on the Web. Access problems include link rot, error pages, login, captchas, etc., and any technical obstacle to search engines bots. This problem causes the no indexation of 9.3% of publications in Google Scholar, 8% in Microsoft Academic, and 6.7% in Semantic Scholar.

Academic search engines also have internal indexing criteria to select the content to be indexed. Google Scholar states that "Content such as news or magazine articles, book reviews, and editorials is not appropriate for Google Scholar" (Google Scholar, 2022), while Microsoft Academic (2021) and Semantic Scholar (2022) do not provide clear information about selection criteria. Because of this, secondary content is not indexed in Google Scholar (6.5%), Semantic Scholar (9.5%), and Microsoft Academic (11.6%). In the case of Google Scholar, the manual inspection showed that this percentage would climb up to 28.3% excluding other causes. In this sense, Google Scholar also claims that "Sites that show [...] bare bibliographic data without abstracts will not be considered for inclusion" (Google Scholar, 2022). Then, publications without a short description about their content (e.g., no abstract) are also excluded in Google Scholar (5.9%) and Microsoft Academic (2.9%). Other document types excluded are Datasets and Posted content, which altogether represent 3.8% of documents in Google Scholar, 4.8% in Microsoft Academic, and 5.6% in Semantic Scholar.

6 | DISCUSSION

This study has attempted to compare different scholarly databases from an original point of view, exploring the reasons behind the no indexation of publications in each of the databases. This new point of view has revealed significant differences between two types of products, bibliographic databases and academic search engines, which build their databases using different methodologies that greatly influence the coverage of publications.

Document typology has been the primary approach in this work. The results have demonstrated this is the most explicative element for detecting coverage differences. More than 80% of the missing documents were explained by their typology. Other variables such discipline or language were less explicative and presented important methodological problems. Nineteen percent of the Crossref documents included a thematic category, and only for journal articles, then the disciplinary analysis would be incomplete and biased. Regarding to language, an initial analysis showed that the proportion of missing documents in all the databases was biased in favor to English-speaking language in a similar proportion, going from 81.4% of English-speaking publications in Google Scholar to the 86.5% of Scilit. Only statistical pairwise differences were found in the case of Google Scholar. Therefore, language analysis was excluded due to the little information that it provided.

6.1 | Reasons for missing publications in classical bibliographic databases

The results have shown bibliographic databases, principally Scilit and Lens, to attain the highest coverage levels relative to Crossref.

The specific causes of non-indexation found (see Table 3) are mainly related to the adaptation of Crossref data to the characteristics of each bibliographic database (internal requirements). While Scilit and Lens ingest Crossref data without remarkable differences, the coverage is higher. However, in those databases where the adaptation process is higher, indexation problems arise. This is the case of Dimensions with book chapters, in which 37.3% of them is not found, being 65.4% of the unmatched documents in that database. This problem was already pointed out by Harzing (2019), who only found one chapter out of 25 in her sample. The reason is that, like search engines, Dimensions indexes 95% of the books from those missing book chapters. This issue is even more striking because the Dimensions core is based on Crossref database, where the book chapters are independently recorded (Hook et al., 2018). The explanation to this lack of book chapter is due to Dimensions does not include book chapters from books labeled *monograph* in Crossref.[4] This would suggest that Dimensions does not consider book chapters as independent publications, because they are conditional to the previous indexation of the book.

These differences between bibliographic databases have been also perceived according to the management of secondary content. While 15.3% of the missing documents in Dimensions match with this category, Scilit and Lens scarcely limit their indexation. This result suggests that these last databases do not have indexation criteria that filter this type of content.

6.2 | Reasons for missing publications in academic search engines

Academic search engines (Google Scholar, Microsoft Academic, and Semantic Scholar) build their databases

[4]Information provided by two Dimensions' reviewers.

23301643, 2024, 1, Downloaded from https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.24839 by Readcube (Labtiva Inc.), Wiley Online Library on [15/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

crawling and harvesting research publications available on the Web, independently of third sources. This might explain their lower coverage of Crossref publications (<90%).

Access problems (either bot exclusion or access problems) constitute the principal cause for missing publications in academic search engines, being 43.2% for Microsoft Academic, 41.9% for Semantic Scholar and 33.3% for Google Scholar. These external factors highlight the important technical limitations of collecting publications from the Web, where metadata are not always accessible or accurate. Perhaps, due to these problems, search engines have stricter indexation criteria. The significant percentage of missing documents in Microsoft Academic (19.3%), Google (16.2%), and Semantic Scholar (15.1%) due to internal requirements report that search engines avoid indexing documents with limited information (e.g., no abstract pages, datasets, comments).

A generalized problem is the coverage of book chapters. Academic search engines do not index these publications properly either. In the case of Google Scholar, there is a technical limitation, a one URL does correspond to only one publication (Delgado López-Cózar et al., 2019). The fact that a book in PDF format can include different independent publications, each of them authored by different authors, is not automatically matched by the indexing algorithm. Chapters are not indexed unless they are independently available with their own URL. Eighty-one percent of the book chapters not indexed in Google Scholar is included in Google Books within the full book, which suggests that Google Books is the main reference for books in Google Scholar and also proves the lack of coordination of these two databases. Microsoft Academic and Semantic Scholar might face similar problems when crawling books. Beyond this technical limitation, the existence of books published as an image instead of text prevents the correct indexation not only of the corresponding chapters but of the references. Lack of commercial agreements with publishers and limitations of book publisher websites might explain the limitation of book chapters indexation.

6.3 | Research implications

These findings have important implications both for the design of scholarly information systems and for research evaluation.

From a technical point of view, the observed differences between bibliographic databases and academic search engines encourage us to recommend using both approaches. This mixed approach could provide a more complete picture about research fields or organizations by combining the scientific literature exploration and the design of accurate information services. The recent case of OpenAlex[5] is a good example of integration of academic search engine data (Microsoft Academic) with external sources (Crossref, Pubmed). This source was tested for inclusion in the study, finding 101,053 (86.6%) records created before 2022. However, in that moment, all the records came from MAG, we accordingly suppose that OpenAlex would not provide more information than the reported by MAG. A recent publication, testing differences between MAG and OpenAlex, showed that, in the early moments, OpenAlex was just a MAG mirror enriched with Crossref's DOIs (Scheidsteger & Haunschild, 2023), being in line with our preliminary results.

For research evaluation, the most problematic result is the incomplete indexation of book chapters. Regardless of the criteria of each database, the absence of a great volume of book chapters in many of the databases undervalues the contribution of researchers and organizations, when these services are used for research evaluation. This problem is especially harmful in research areas with a high production of book and book chapters, such as social sciences and humanities (Huang & Chang, 2008).

Methodologically, this study has evidenced that the size and coverage of databases should be interpreted according to the reference sample used in the analysis, because this always introduces a selection bias. The most illustrative example in our case is Google Scholar, accounted as the largest scholarly information service (Gusenbauer, 2019; Martín-Martín et al., 2021), but with a lower coverage of items that are deposited in Crossref. This result does not invalidate Google Scholar as the largest scholarly information service, but illustrates that there is a considerable amount of scientific literature that is not indexed in Google Scholar. Previous studies already warned on this fact (Adriaanse & Rensleigh, 2013; Bar-Ilan, 2010; Giustini & Boulos, 2013; Martín-Martín & López-Cózar, 2021). This consideration leads us to a second criticism to coverage studies: it should take more into account the quality and value of the indexed documents than the mere number of publications. Thus, for example, the fact that Dimensions or Google Scholar cover fewer publications from Crossref than Scilit or Lens should not be seen as a weakness, but as a sign that these services have stricter indexation criteria, selecting publications with a rich scientific content (e.g., journal articles, book chapters) and filtering out scarcely informative items (e.g., indexes, announcements, front covers, prefaces, glossaries). This fact has important implications in the appreciation of scholarly databases because if a

[5] https://openalex.org.

23301643, 2024, 1, Downloaded from https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.24839 by Readcube (Labtiva Inc.), Wiley Online Library on [15/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

database does not filter and select content, then it does not add value and therefore its use is less attractive. Precisely, because the lack of content processing would cause noise in the retrieval of documents and inflated coverage.

The comparative study of bibliographic sources always deals with data access problems that make it difficult to value the performance of these services. These problems are more evident in the case of commercial platforms, some of which impede their data be used for research purposes or condition the data usage on the approval of a research project proposal, considering aspects beyond the technical use of their servers and downloading services. We understand that these policies constrain the development of research focused on describing how these platforms operate. This is undesirable as many of these platforms take their data from open sources such as Crossref, PubMed, or MAG. This is added to the fact that information about content selection is sometimes limited (e.g., Dimensions, Google Scholar, Semantic Scholar) or even absent (e.g., Scilit, Lens). This makes difficult a more detailed discussion about to what extent indexation criteria determine the non-coverage of publications.

6.4 | Limitations

A third-party study is determined by the coverage limitations of the reference sample. In our case, Crossref only includes publications from partner publishers, leaving aside some conference proceedings and local journals (see footnote 2). These publications are not curated, being able to include non-strictly research materials. The identification of *secondary content* indicates that publishers deposit any type of material, regardless of the scholarly content. This issue underscores a second limitation of Crossref. Document typologies are not precise because publishers may confuse or misattribute typologies. We have encountered this issue with Dimensions, where book chapters from books with *monograph* type are not indexed. Similarly, manual inspection of Google Scholar revealed that 32.8% of missing journal articles fell into different categories.

Another problem could stem from the extraction process. In the case of Dimensions, Semantic Scholar, and Microsoft Academic, specific R packages were used to query these services (dimensionsR, microdemic, and semscholar). Our experience shows us that all these packages present some type of bug or error, which leads us to directly query the API in some cases. This problem could have caused some type of loss of information.

Searching by DOIs introduces the risk that this identifier could not be assigned to the document in the searched database (Van Eck et al., 2018). This issue has been clear in Microsoft Academic. To mitigate this risk, searches by title were conducted in those cases in which a reliable endpoint was not available. The slight improvement in Google Scholar has shown that this form of search requires considerably more effort than the reward that is received.

In the specific case of Google Scholar, we found problems with the results page. The first one is when we search by title, documents with a very short title and common words did not produce exact matching, and several items were showed. This made very difficult and time-consuming to identify the correct document. Another problem was the false positives, when a DOI query retrieves a wrong document because it mentions that DOI in the abstract (e.g., retractions). These retrieval problems lead us to point out that limited search functionalities would influence on the matching of documents, distorting the real coverage, as could be the case of Google Scholar (Boeker et al., 2013; Gusenbauer & Haddaway, 2020).

7 | CONCLUSIONS

The results of this study allow us to conclude that, from the Crossref point of view, there are remarkable coverage differences between scholarly databases. These differences are mainly due to methodological approaches used by each database to build their databases. The proportion of missing documents has evidenced that bibliographic databases, such as Scilit (0.2%) and Lens (0.1%), almost exactly reproduce the content of Crossref. However, academic search engines, such as Microsoft Academic (12%) and Semantic Scholar (10%), showed important absence of records. Dimensions (7.6%) and Google Scholar (9.8%) stand in an intermediate position. However, these coverage differences should be critically considered because a high coverage of Crossref records also implies low filtering levels of scholarly publications, causing noise in the retrieval and poor content curation.

The cause of these disparities is principally due to the management of specific document types. Bibliographic databases experience more problems covering journal articles (>75%), and search engines find limitations both in journal articles (≈40%) and book chapters (≈40%).

The reasons of this non-indexation of documents are different according to the type of scholarly product. For bibliographic databases, such as Dimensions, is due to internal selection criteria that index full books instead book chapters (65%) and exclude secondary content (15%). In the event of academic search engines, there are important external limitations (web accessibility, robot

23301643, 2024, 1, Downloaded from https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.24839 by Readcube (Labtiva Inc.), Wiley Online Library on [15/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

restrictions) that prevent the indexation of research documents (39.2%–46%), and internal requirements that exclude secondary content (6.5%–11.6%).

This work represents an advance in the study of bibliographic databases coverage, by introducing the reference sample (third party) method, and by considering free-access bibliographic databases and academic search engines. The results obtained have made it possible to know accurately the reasons for the non-indexing of documents, identifying specific motives according to the type of database (classical databases or academic search engines). These results are helpful to meta-researchers, when learning about the characteristics of the databases used in bibliometric studies, as well as to librarians and practitioners who need to use scholarly databases to assist researchers or carry out training tasks. Likewise, it uncovers the need for publishers to properly update their websites and reach specific agreements with academic search engines to be correctly indexed by these products, which are called to coexist with the classic databases.

ACKNOWLEDGMENTS

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. The authors thank Digital Science for access to the Dimensions API.

CONFLICT OF INTEREST STATEMENT

José Luis Ortega is member of the Scilit advisory board, which made possible an internal endpoint to that database. In addition to the two anonymous reviewers, two employees from Digital Science (Dimensions) participated in the review process at the request of the Editor-in-Chief. This should not be interpreted as an endorsement of neither the methods nor the results.

ORCID

José Luis Ortega https://orcid.org/0000-0001-9857-1511

REFERENCES

Adriaanse, L. S., & Rensleigh, C. (2013). *Web of Science, Scopus and Google Scholar: A content comprehensiveness comparison*. The Electronic Library.

Bakkalbasi, N., Bauer, K., Glover, J., & Wang, L. (2006). Three options for citation tracking: Google Scholar, Scopus and Web of Science. *Biomedical Digital Libraries*, *3*(1), 1–8.

Bar-Ilan, J. (2010). Citations to the "Introduction to informetrics" indexed by WOS, Scopus and Google Scholar. *Scientometrics*, *82*(3), 495–506.

Boeker, M., Vach, W., & Motschall, E. (2013). Google Scholar as replacement for systematic literature searches: Good relative recall and precision are not enough. *BMC Medical Research Methodology*, *13*(1), 1–12.

Borgman, C. L., & Furner, J. (2002). Scholarly communication and bibliometrics. *Annual Review of Information Science and Technology*, *36*(1), 2–72.

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/

Bramer, W. M., Rethlefsen, M. L., Kleijnen, J., & Franco, O. H. (2017). Optimal database combinations for literature searches in systematic reviews: A prospective exploratory study. *Systematic Reviews*, *6*(1), 1–12.

Delgado López-Cózar, E., Orduña-Malea, E., & Martín-Martín, A. (2019). Google Scholar as a data source for research assessment. In *Springer handbook of science and technology indicators* (pp. 95–127). Springer.

Dimensions Plus. (2019). What exactly is covered in the "Publications" in Dimensions? https://plus.dimensions.ai/support/solutions/articles/23000018859-what-exactly-is-covered-in-the-publications-in-dimensions

Fiala, D. (2011). Mining citation information from CiteSeer data. *Scientometrics*, *86*(3), 553–562.

Giustini, D., & Boulos, M. N. K. (2013). Google Scholar is not enough to be used alone for systematic reviews. *Online Journal of Public Health Informatics*, *5*(2), 214.

Google Scholar. (2022). Google Scholar Help: Inclusion guidelines for webmasters. Retrieved from https://scholar.google.com/intl/en/scholar/inclusion.html#content

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*, 19.

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. https://doi.org/10.1007/s11192-018-2958-5

Gusenbauer, M., & Haddaway, N. R. (2020). Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. *Research Synthesis Methods*, *11*(2), 181–217.

Haley, M. R. (2014). Ranking top economics and finance journals using Microsoft academic search versus Google scholar: How does the new publish or perish option compare? *Journal of the Association for Information Science and Technology*, *65*(5), 1079–1084.

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.

Harzing, A. W., & Alakangas, S. (2017). Microsoft Academic: Is the phoenix getting wings? *Scientometrics*, *110*(1), 371–383.

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.

Herzog, C., Hook, D., & Konkiel, S. (2020). Dimensions: Bringing down barriers between scientometricians and data. *Quantitative Science Studies*, *1*(1), 387–395.

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

Huang, M. H., & Chang, Y. W. (2008). Characteristics of research output in social sciences and humanities: From a research

23301643, 2024, 1, Downloaded from https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.24839 by Readcube (Labtiva Inc.), Wiley Online Library on [15/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

evaluation perspective. *Journal of the American Society for Information Science and Technology*, *59*(11), 1819–1828.

Hug, S. E., & Brändle, M. P. (2017). The coverage of Microsoft Academic: Analyzing the publication output of a university. *Scientometrics*, *113*(3), 1551–1571. https://doi.org/10.1007/s11192-017-2535-3

Jacso, P. (2005). As we may search—Comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases. *Current Science*, *89*(9), 1537–1547.

Khabsa, M., & Giles, C. L. (2014). The number of scholarly documents on the public web. *PLoS One*, *9*(5), e93949.

Kousha, K., Thelwall, M., & Rezaie, S. (2011). Assessing the citation impact of books: The role of Google Books, Google Scholar, and Scopus. *Journal of the American Society for Information Science and Technology*, *62*(11), 2147–2164.

Levine-Clark, M., & Gil, E. L. (2008). A comparative citation analysis of Web of Science, Scopus, and Google Scholar. *Journal of Business & Finance Librarianship*, *14*(1), 32–46.

Martín-Martín, A., & Delgado López-Cózar, E. D. (2021). Large coverage fluctuations in Google Scholar: A case study. arXiv preprint arXiv:2102.07571.

Martín-Martín, A., Orduna-Malea, E., & Delgado López-Cózar, E. (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., & Delgado López-Cózar, 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.

Martín-Martín, A., Thelwall, M., Orduna-Malea, E., & Delgado López-Cózar, E. (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. https://doi.org/10.1007/s11192-020-03690-4

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.

Microsoft Academic. (2021). Frequently Asked Questions. Retrieved from https://web.archive.org/web/20211223093934/https://academic.microsoft.com/faq

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: A comparative analysis. *Scientometrics*, *106*(1), 213–228.

Orduña-Malea, E., Ayllón, J. M., Martín-Martín, A., & Delgado López-Cózar, E. (2015). Methods for estimating the size of Google Scholar. *Scientometrics*, *104*(3), 931–949.

Orduña-Malea, E., & Delgado-López-Cózar, E. (2018). Dimensions: re-discovering the ecosystem of scientific information. *El Profesional de la Información*, *27*(2), 1699–2407.

Orduña-Malea, E., Martín-Martín, A., & Delgado López-Cózar, E. D. (2017). Google Scholar as a source for scholarly evaluation: A bibliographic review of database errors. *Revista Española de Documentación Científica*, *40*(4), e185.

Ortega, J. L. (2014). *Academic search engines: A quantitative outlook*. Chandos (Elsevier).

Ortega, J. L. (2021). El movimiento Open Citations y sus implicaciones en la transformación de la evaluación científica. *Arbor*, *197*(799), a592.

Ortega, J. L., & Aguillo, I. F. (2014). Microsoft academic search and Google scholar citations: Comparative analysis of author profiles. *Journal of the Association for Information Science and Technology*, *65*(6), 1149–1156.

Penfold, R. (2020). Using the Lens database for staff publications. *Journal of the Medical Library Association*, *108*(2), 341–344.

Purnell, P. J. (2022). The prevalence and impact of university affiliation discrepancies between four bibliographic databases—Scopus, Web of Science, Dimensions, and Microsoft Academic. *Quantitative Science Studies*, *3*(1), 99–121. https://doi.org/10.1162/qss_a_00175

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? *El Profesional de la Información*, *32*(2), e320209. https://doi.org/10.3145/epi.2023.mar.09

Semantic Scholar. (2022). Frequently Asked Questions. Retrieved from https://www.semanticscholar.org/faq#content-types

Singh, V. K., Singh, P., Karmakar, M., Leta, J., & Mayr, P. (2021). The journal coverage of Web of Science, Scopus and Dimensions: A comparative analysis. *Scientometrics*, *126*(6), 5113–5142.

Thelwall, M. (2018). Dimensions: A competitor to Scopus and the Web of Science? *Journal of Informetrics*, *12*(2), 430–435. https://doi.org/10.1016/j.joi.2018.03.006

Van Eck, N. J., Waltman, L., Larivière, V., & Sugimoto, C. (2018). Crossref as a new source of citation data: A comparison with Web of Science and Scopus. *CWTS Blog*. Retrieved from https://www.cwts.nl/blog?article=n-r2s234&sthash.lInLf4Uz.mjjo

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. https://doi.org/10.1162/qss_a_00112

Wang, K., Shen, Z., Huang, C., Wu, C. H., Dong, Y., & Kanakia, A. (2020). Microsoft academic graph: When experts are not enough. *Quantitative Science Studies*, *1*(1), 396–413. https://doi.org/10.1162/qss_a_00021

Wang, K., Shen, Z., Huang, C., Wu, C. H., Eide, D., Dong, Y., Qiang, J., Kanakia, A., Chen, A., & Rogahn, R. (2019). A review of microsoft academic services for science of science studies. *Frontiers in Big Data*, *2*, 45. https://doi.org/10.3389/fdata.2019.00045

How to cite this article: 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. https://doi.org/10.1002/asi.24839

23301643, 2024, 1, Downloaded from https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.24839 by Readcube (Labtiva Inc.), Wiley Online Library on [15/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

APPENDIX A

TABLE A1 Distribution of indexed publications in the Crossref sample by document type and the proportion of missing documents in the remaining databases

Type_doc	Crossref	Crossref %	Dimensions	Dimensions %	Google Scholar	Google Scholar %	Lens	Lens %	Microsoft Academic	Microsoft Academic %	Scilit	Scilit %	Semantic Scholar	Semantic Scholar %
Addendum	1	0.0%												
Book	1263	1.1%	246	2.8%	190	1.7%			410	2.9%	1	0.4%	329	2.8%
Book-chapter	15,120	13.0%	5843	65.9%	4449	39.1%	3	12.5%	5313	38.0%	15	5.8%	4948	41.7%
Book-set	1	0.0%	1	0.0%										
Component	270	0.2%	194	2.2%	231	2.0%			198	1.4%		0.0%	175	1.5%
Correction	191	0.2%			11	0.1%			74	0.5%		0.0%	11	0.1%
Corrigendum	5	0.0%							2	0.0%		0.0%		
Dataset	349	0.3%	348	3.9%	304	2.7%			288	2.1%	10	3.9%	239	2.0%
Dissertation	1	0.0%												
Erratum	128	0.1%	1	0.0%	3	0.0%			30	0.2%		0.0%	12	0.1%
Journal	2	0.0%	1	0.0%									1	0.0%
journal-article	86,908	74.5%	1086	12.3%	4766	41.9%	18	75.0%	5846	41.8%	173	66.8%	4676	39.4%
Journal-issue	205	0.2%	189	2.1%	175	1.5%			198	1.4%	24	9.3%	69	0.6%
Monograph	520	0.4%			47	0.4%			192	1.4%	2	0.8%	126	1.1%
Other	227	0.2%	104	1.2%	156	1.4%			172	1.2%			172	1.4%
Peer-review	39	0.0%	39	0.4%	19	0.2%			28	0.2%	14	5.4%	25	0.2%
Posted-content	935	0.8%	533	6.0%	344	3.0%			619	4.4%	4	1.5%	569	4.8%
Proceedings	32	0.0%	31	0.3%	12	0.1%			23	0.2%			16	0.1%
Proceedings-article	10,111	8.7%	123	1.4%	547	4.8%	3	12.5%	404	2.9%	15	5.8%	321	2.7%
Proceedings-series	2	0.0%	2	0.0%					1	0.0%			1	0.0%
Reference-book	48	0.0%	25	0.3%	23	0.2%			28	0.2%			28	0.2%
Reference-entry	168	0.1%			85	0.7%			134	1.0%			130	1.1%

(Continues)

23301643, 2024, 1, Downloaded from https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.24839 by ReadCube (Labtiva Inc.), Wiley Online Library on [15/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE A1 (Continued)

Type_doc	Crossref	Crossref %	Dimensions	Dimensions %	Google Scholar	Google Scholar %	Lens	Lens %	Microsoft Academic	Microsoft Academic %	Scilit	Scilit %	Semantic Scholar	Semantic Scholar %
Report	99	0.1%	91	1.0%	14	0.1%			22	0.2%	1	0.4%	23	0.2%
Report-series	5	0.0%	3	0.0%					1	0.0%			1	0.0%
Retraction	17	0.0%							2	0.0%				
Total	116,647	100.0%	8860	100.0%	11,376	100.0%	24	100.0%	13,985	100.0%	259	100.0%	11,872	100.0%

23301643, 2024, 1, Downloaded from https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.24839 by ReadCube (Labtiva Inc.), Wiley Online Library on [15/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License