RESEARCH ARTICLE

Completeness degree of publication metadata in eight free-access scholarly databases

Lorena Delgado-Quirós[1,2] **and José Luis Ortega**[1,2]

[1]Institute for Advanced Social Studies (IESA-CSIC), Córdoba, Spain
[2]Joint Research Unit Knowledge Transfer and Innovation (UCO-CSIC), Córdoba, Spain

Keywords: academic search engine, completeness degree, metadata quality, open access, scholarly bibliographic databases, third-party databases

ABSTRACT

The main objective of this study is to compare the amount of metadata and the completeness degree of research publications in new academic databases. Using a quantitative approach, we selected a random Crossref sample of more than 115,000 records, which was then searched in seven databases (Dimensions, Google Scholar, Microsoft Academic, OpenAlex, Scilit, Semantic Scholar, and The Lens). Seven characteristics were analyzed (abstract, access, bibliographic info, document type, publication date, language, and identifiers), to observe fields that describe this information, the completeness rate of these fields, and the agreement among databases. The results show that academic search engines (Google Scholar, Microsoft Academic, and Semantic Scholar) gather less information and have a low degree of completeness. Conversely, third-party databases (Dimensions, OpenAlex, Scilit, and The Lens) have more metadata quality and a higher completeness rate. We conclude that academic search engines lack the ability to retrieve reliable descriptive data by crawling the web, and the main problem of third-party databases is the loss of information derived from integrating different sources.

1. INTRODUCTION

The recent proliferation of bibliographic scholarly databases has stimulated interest in these new platforms, given their possibilities of finding scientific literature and providing different bibliometric indicators. This attention has focused on testing the performance of these new systems in relation to traditional products, such as citation indexes (e.g., Web of Science (WoS), Scopus) and academic search engines (e.g., Google Scholar, Microsoft Academic). These new products could be defined as hybrid databases because they share characteristics with the former. On the one hand, these platforms also extract and process citations for computing *ad hoc* bibliometric indicators as classical citation indexes. On the other hand, they are similar to search engines because they opt for a free-access model in which users require no subscription fee to search and retrieve documents. Moreover, in some cases, they provide open data through REST APIs or dump files.

However, these hybrid products have some particularities that make them different. Most importantly, they are fed by third-party sources. The appearance of Crossref as a repository of publishers' metadata, the availability of APIs and dump files from academic search engines (e.g., Microsoft Academic, Semantic Scholar), and the possibility of reusing other bibliographic databases (e.g., PubMed, Directory of Open Access Journals (DOAJ), repositories) have

an open access journal

Check for updates

Citation: 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. https://doi.org/10.1162/qss_a_00286

DOI: https://doi.org/10.1162/qss_a_00286

Peer Review: https://www.webofscience.com/api/gateway/wos/peer-review/10.1162/qss_a_00286

Received: 18 May 2023
Accepted: 10 December 2023

Corresponding Author:
José Luis Ortega
jortega@iesa.csic.es

Handling Editor:
Vincent Larivière

The MIT Press

facilitated the emergence of these bibliographic products that have quickly and inexpensively covered much of the scientific literature.

However, this multiple and varied availability of bibliographic data also presents a challenge for these new platforms because the integration of data from different sources requires intense data processing to prevent duplicate records, filter nonscholarly materials, and manage different versions of the same document. All of these actions are addressed to achieve optimal metadata quality, which could be defined as the quantity of information that describes an entity in an accurate and reliable form. The integration of internal and external descriptions determines this quality to a large extent.

For this reason, examining the quality of publication metadata in the new scholarly databases allows us to appreciate the extent to which these processing efforts are accomplished and to value the suitability and reliability of these search tools to provide rich information about scientific literature. This study aims to explore the metadata publication quality of these new databases to obtain a global picture about the richness of the information provided by each platform.

2. LITERATURE REVIEW

Many studies have focused on the evaluation of the performance of these new academic databases, comparing the coverage and overlap of records (Gusenbauer, 2019; Martín-Martín, Thelwall et al., 2021; van Eck, Waltman et al., 2018; Visser, van Eck, & Waltman, 2021). This quantitative procedure is excessively centered on the size of each platform and overlooks the amount and quality of the content included in each database. In this sense, some articles have described the metadata quality of specific sources as a way of informing about the richness and limitations of those sources. Hendricks, Tkaczyk et al. (2020) described how the Crossref database functions and analyzed the completeness of its metadata. Similar papers were published describing Semantic Scholar (Wade, 2022), The Lens (Jefferson, Koellhofer et al., 2019), Dimensions (Herzog, Hook, & Konkiel, 2020), and Microsoft Academic (Wang, Shen et al., 2020). Many of these studies were descriptive reviews written by their employees with no critical discussion of the quality of the data.

In other cases, the coverage of certain elements or entities in different scholarly databases was studied to test their performance in processing specific information. Hug and Brändle (2017) analyzed in detail the coverage of Microsoft Academic, finding important problems in the assigning of author and publication data in comparison to WoS and Scopus. Ranjbar-Sahraei and van Eck (2018) also tested the problems of Microsoft linking papers with organizations. Guerrero-Bote, Chinchilla-Rodríguez et al. (2021) compared affiliation information between Scopus and Dimensions and found that close to half of all documents in Dimensions were not associated with any country. Purnell (2022) evaluated affiliation discrepancies in four scholarly databases, finding that the larger the database, the greater the disambiguation problems. Kramer and de Jonge (2022) analyzed the information about funders included by Crossref, The Lens, WoS, Scopus, and Dimensions, uncovering important differences when they came to extract and process that information. Lutai and Lyubushko (2022) also analyzed the coverage in six databases, detecting discrepancies and similarities in the identification and indexation of Russian authors.

As for publications, some studies have explored the amount of information and quality of this key entity in scholarly databases. Herrmannova and Knoth (2016) tested the reliability of the publication date in Microsoft Academic Graph, finding that 88% of cases showed a correct date. Liu, Hu, and Tang (2018) detected that approximately 20% of WoS publications have no

information in the address field. Basson, Simard et al. (2022) showed that the proportion of open access documents in Dimensions is higher than in WoS because the former indexes more publications from Asian and Latin American countries. Other studies have examined errors and inconsistencies in different academic databases to test their suitability for bibliometric studies or for bibliographic searches only. Thus, some articles have analyzed duplicate records management in Scopus (Valderrama-Zurián, Aguilar-Moya et al., 2015) and WoS (Franceschini, Maisano, & Mastrogiacomo, 2016).

Metadata quality is a concept that has emerged recently as a result of the abundance of available data sets and sources. Bruce and Hillman (2004) defined seven characteristics that should be considered when assessing the quality of metadata: completeness, accuracy, provenance, conformance to expectations, logical consistency and coherence, timeliness, and accessibility. Subsequently, Ochoa and Duval (2009) proposed varied metrics to automatically assess these features. However, many of these studies have been applied to digital libraries or repositories (Kubler, Robert et al., 2018; Tani, Candela, & Castelli, 2013), with an important gap in the analysis of scholarly databases.

3. OBJECTIVES

The main objective of this study is to compare the metadata quality of research publications in the new academic databases, using a quantitative approach, to describe the advantages and limitations of these scholarly platforms in their provision of bibliographic information for analytical studies and secondary products. To this end, seven of these new bibliographic databases were analyzed for their coverage of a random sample of publications from Crossref. The following research questions were formulated:

- Is it possible to quantitatively compare the metadata content of different databases using a third source (i.e., Crossref)? What advantages and limitations could this procedure report?
- Do similarities or discrepancies among databases allow us to delimit different models of databases, with their advantages and limitations?
- Which databases provide the most metadata and have a higher completeness rate?

4. METHODS

4.1. Source Selection Criteria

This comparative approach entails the selection of equitable samples to benchmark bibliographic databases and observe what information about publications is indexed (e.g., bibliographic information, publication dates, identifiers, and metrics). Seven bibliographic databases were considered for the study: Crossref, Dimensions, The Lens, Microsoft Academic, OpenAlex, Scilit, and Semantic Scholar. Two requisites were considered for selecting these sources:

- They had to be freely accessible through the web: a free-subscription search interface.
- They had to provide metrics for research evaluation.

4.2. Sample Selection and Extraction

Crossref was selected as a control sample for several reasons. The first was an operational question. Crossref is a publishers' consortium that assigns the Digital Object Identifier

(DOI), the most extended persistent identifier of research publications in the publishing system. Although coverage is limited to publisher members (Visser et al., 2021), its use is justified because all these platforms allow publications to be queried by DOIs, thereby facilitating rapid and exact matching. The second reason is related to methodological issues: Crossref enables the extraction of random samples of documents (https://api.crossref.org/works?sample=100). The representativeness of the sample is thus reinforced because it avoids the influence of ranking algorithms, filters, or matching procedures that could distort the quality of the sample. A third motive is that publishers can request a DOI for any published material, regardless of typology, discipline, or language. The Crossref database therefore has no inclusion criteria that could limit the coverage of certain types of documents (e.g., indexes, acknowledgements, or front covers). This nonselective criterion would lead us to clearly appreciate the inclusion policies of the different bibliographic platforms. Finally, Crossref is fed by publishers that deposit metadata about their publications and could be considered the most authoritative source of the reliability and accuracy of their own publications.

4.3. Sources Description

This section describes each source analyzed:

- **Dimensions**: Created in 2018 by Digital Science, this database is supported by external products, Crossref being the main source (Hook, Porter, & Herzog, 2018). It gathers more than 138 million publications, in addition to patents (154 million), data sets (12 million), and grants (7 million).
- **Google Scholar**: One of the most important academic search engines due to its estimated size (389 million) and age (2004) (Gusenbauer, 2019), it obtains data directly from the web. Special agreements with publishers allow it to extract citations and content information from paywalled journals. Its search interface can also access books (Google Books) and patents (Google Patents).
- **OpenAlex**: This is the newest service (2022). It was created by OurResearch, a nonprofit organization that recovered the defunct Microsoft Academic Graph (203 million) to implement a new open product. The core of OpenAlex was then set up by Microsoft Academic Graph, with the addition of data from other open sources such as Crossref, PubMed, and ORCID. OpenAlex now indexes 240 million publications.
- **Microsoft Academic**: The last version of this search engine functioned between 2016 and 2021. Like Google Scholar, it also crawled the web extracting metadata from scholarly publications, reaching 260 million publications. Part of its database was openly released (Microsoft Academic Graph), which has enabled reuse by other bibliographic products.
- **Scilit**: Active since 2016, this database was created by the publisher MDPI as a way to compete in the scholarly database market. It indexes 159 million scholarly publications, taken mainly from Crossref and PubMed.
- **Semantic Scholar**: This search engine was launched in 2015 by the Allen Institute for Artificial Intelligence. Although it uses crawlers to extract information from the web, in 2018, it agreed to include Microsoft Academic Graph as a primary source (Boyle, 2018). It now encompasses almost 214 million research papers.
- **The Lens**: This database was developed by Cambia, a nonprofit organization, in 2000. It initially started as a patent database, but in 2018 incorporated scholarly publications from Crossref, PubMed, and Microsoft Academic. It now provides more than 225 million scholarly works.

4.4. Data Retrieving

A sample of 116,648 DOIs was randomly extracted from Crossref in August 2020 and July 2021, the only limitation being that documents were to have been published between 2014 and 2018. This time window was selected so that publications could accrue a significant number of citations and other metrics. The sample was generated by sending 1,200 automatic requests to https://api.crossref.org/works?sample=100. Duplicate records produced by this random process were removed to obtain the final list. The resulting distribution by document type largely coincides with the entire database (Hendricks et al., 2020), thereby reinforcing the reliability of the sample. Hendricks et al. (2020) adapt some categories, in which Book presumably includes book and book-chapter, and Preprints might be posted-content (Table 1).

Next, this control sample was queried on each platform to match the records and extract all the information about each publication. This task was performed in July 2021, except for Scilit and OpenAlex. In the case of Scilit, data were retrieved in December 2022 because a new public API, with more information, was launched in June 2022. OpenAlex was added to the study in January 2023 because of its novelty as an open bibliographic source. The extraction process in each platform is described in detail:

- **Dimensions**: This database was accessed through the API (https://app.dimensions.ai/dsl/v2). An R package (dimensionsR[1]) was used to extract the data. The results were downloaded in the JSON format because dimensionsR caused problems in the transformation of JSON outputs to CSV format.
- **Google Scholar**: Web scraping was used to automatically query each DOI in the search box because Google Scholar does not facilitate access to its data. The RSelenium[2] R package was used to emulate a browser session and avoid antirobot actions (i.e., CAPTCHAs). Because some DOIs could not be indexed (Martín-Martín, Orduna-Malea et al., 2018), a title search with the query "allintitle:title" was used to complete the results.
- **OpenAlex**: This bibliographic repository was accessed through its public API (https://api.openalex.org/). A Python routine was written to extract and process the data.
- **Microsoft Academic**: Coverage of this service was obtained by several methods. Firstly, 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. Microdemic[3], an R package, was used to query the API. However, the low indexation of DOIs (37.1%) and their case sensitivity forced us to download the entire table of publications available in Zenodo (https://zenodo.org/record/2628216) and locally match them with the sample, using DOIs and titles.
- **Scilit**: This platform was accessed via a public API (https://app.scilit.net/api/v1/). Because access must be via a POST method, a Python script was designed to extract the data.
- **Semantic Scholar**: This database provides a public API (https://api.semanticscholar.org/v1). The semscholar[4] R package was used to extract the data. However, the API was directly queried subsequently to detect any problems in the retrieval process. A Python script was written.

[1] https://github.com/massimoaria/dimensionsR.
[2] https://docs.ropensci.org/RSelenium/.
[3] https://docs.ropensci.org/microdemic/.
[4] https://github.com/njahn82/semscholar.

Table 1. Document type distribution in the Crossref random sample and the total coverage in June 2019 (Hendricks et al., 2020). Only the first 10 most frequent categories are shown

Crossref random sample			Hendricks et al. (2020)		
Document type	**Publications**	**Percentage**	**Document type**	**Publications**	**Percentage**
journal-article	87,091	74.7	Journal	77,738,314	73.3
book-chapter	15,119	13.0	Book	13,707,631	12.9
proceedings-article	10,111	8.7	Conference paper & proceedings	5,854,822	5.5
book	1,260	1.1	Component	3,771,565	3.6
posted-content	935	0.8	Data set	1,783,953	1.7
monograph	520	0.4	Reference book	1,079,324	1.0
data set	349	0.3	Report	615,879	0.6
component	270	0.2	Monograph	421,987	0.4
other	208	0.2	Other	326,437	0.3
report	99	0.1	Preprints	119,379	0.1
Total	**116,648**		**Total**	**106,012,923**	

- **The Lens**: After a formal request, this service provided temporary access to its API (https://api.lens.org/scholarly/search). In this case, an R script was written to directly extract the data. However, some relevant fields (i.e., abstract, source_urls, funders) for this study were not properly retrieved, due to technical reasons, in July 2021. We then decided to extract a small sample of 5,000 records directly from the main search page (https://www.lens.org/lens/) to make up for this limitation in January 2023. From that request, 4,996 records were successfully retrieved.

This study follows a qualitative-quantitative approach. On the one hand, large data samples from different sources were extracted for subsequent comparison of the data completeness degree in several fields (quantitative approach). On the other, we analyzed the quality features of these data: the richness of fields to describe an entity (more/less information about dates, identifiers, accesses, etc.), the reliability of the data (agreement between sources), or the degree of data processing data (e.g., document type classification). API documentation about each database was reviewed to discover the metadata available about publications (qualitative approach). Table 2 lists the seven sources analyzed, the total amount of publications retrieved, and the link with the description about the provided fields.

According to Bruce and Hillman (2004), this study will focus on the evaluation of four characteristics of metadata quality:

- Completeness: the quantity of fields and degree of filling for an entity.
- Accuracy: proportion of data integrity with respect to the original data or sources.
- Provenance: description of primary data sources.
- Logical consistency and coherence: the degree of resemblance in the description and classification of objects (e.g., language, document type).

Table 2. Web source with information about publication metadata in each database

Database	Publications	Information about publication metadata
Crossref	116,592	https://github.com/CrossRef/rest-api-doc/blob/master/api_format.md
Dimensions	105,062	https://docs.dimensions.ai/dsl/datasource-publications.html#publications-authors-long-desc
Google Scholar	101,691	https://scholar.google.com/
Microsoft Academic	96,336	https://web.archive.org/web/20230329104454/https://learn.microsoft.com/en-us/academic-services/graph/reference-data-schema
The Lens	116,337 (4,996)	https://docs.api.lens.org/response-scholar.html
OpenAlex	115,881	https://docs.openalex.org/
Scilit	113,422	No public information
Semantic Scholar	92,314	https://api.semanticscholar.org/api-docs/graph

5. RESULTS

This study describes the amount and quality of metadata associated with the description of research publications indexed in these databases. Publications are central to the publishing ecosystem and are therefore the main asset of a bibliographic database. A clear and complete description of their elements and characteristics improves the identification and retrieval of these items, and their connection with other entities. As a result, publications are the entity with most fields, ranging from 38 fields in Crossref to 18 in Semantic Scholar. Next, we analyze the fields used by each database to describe the main characteristics of a publication.

5.1. Abstract

This is an important access point to publication content because it provides a summary of the research. All the databases analyzed index this element. For Microsoft Academic, the table with this information (*PaperAbstractsInvertedIndex*) is not yet available. However, early studies detected a coverage of 58% (Färber & Ao, 2022). Google Scholar does not index the publication abstract, but it does extract parts of the document text (Google Scholar, 2023).

Table 3 shows the number and proportion of publications with an abstract. Google Scholar indexes the most articles with a summary (91.66%), although some are merely an extract of the

Table 3. Proportion of publications with abstract in each database. Margin of error at 99%

Database	Field	Publications	Percentage
Crossref	*abstract*	15,927	13.66 (±0.26)
Dimensions	*abstract*	73,145	69.62 (±0.37)
Google Scholar		93,215	91.66 (±0.22)
The Lens	*abstract*	3,133	62.7 (±1.76)
Scilit	*abstract*	57,300	50.52 (±0.38)
Semantic Scholar	*abstract*	50,263	54.45 (±0.42)
OpenAlex	*abstract_inverted_index*	73,899	63.77 (±0.36)

text. Aside from this, Dimensions indexes most articles with their abstract (69.6%), followed by OpenAlex (63.8%). Conversely, Crossref has fewer publications with an abstract (13.7%). This last percentage is lower than that reported by Waltman, Kramer et al. (2020) (21%), perhaps because our study also gathers other materials, such as book chapters and conference papers, which do not always include a formal abstract. This low percentage of abstracts in Crossref shows that this information is not usually provided by publishers. Indexation services therefore need to process documents to obtain this data, which would explain the overall low availability of abstracts in free-access databases, in particular in Scilit (50.5%) and Semantic Scholar (54.45%). Only OpenAlex and The Lens (χ^2 = 2.2881, *p*-value = 0.1304) show no statistical differences, perhaps due to the small sample size of The Lens or because both databases use Microsoft Academic as a source.

5.2. Access

Today, a positive feature of scholarly databases is that they provide some type of access to original publications. Widespread electronic publishing allows for the provision of links to different venues where the document is partially or fully hosted. All the databases include external links to the original publication. Microsoft Academic and Crossref have no specific field for open access publications. In March 2023, OpenAlex changed some fields about access: *Landing_page_url* is designated as an external link, but only includes DOI links (99.8%), and *oa_url* indexes external links to open access versions only. We therefore concluded that OpenAlex includes external links for open access publications only. This is also the case for Dimensions, which only indexes external links (*linkout*) for open_access (*open_access*) articles.

Figure 1 and Table 4 depict the percentage of publications with external links to the original source and information about whether or not they are open access per database. Google Scholar (97.1%) includes the most external links (97.1%), followed by The Lens (82.9%), Microsoft Academic (80.8%), and Crossref (79%). It is evident that the academic search

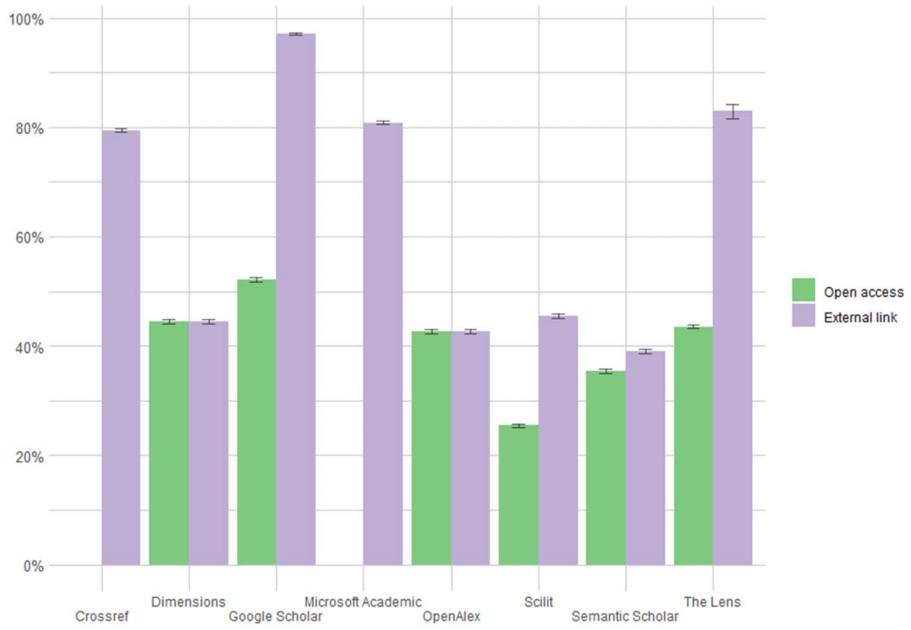

Figure 1. Proportion of bibliographic records with information about open access and external links.

Table 4. Fields, publications, and percentage of publications with external links and information about open access by database. Margin of error at 99%

	External links			Open access		
Database	**Field**	**Publications**	**Percentage**	**Field**	**Publications**	**Percentage**
Crossref	*Link*	92,561	79 (±0.3)			
Dimensions	*Linkout*	46,732	44.5 (±0.4)	*open_access*	46,729	44.5 (±0.4)
Google Scholar		98,714	97.1 (±0.1)		53,034	52.2 (±0.4)
Microsoft Academic	*PaperURL*	77,877	80.8 (±0.3)			
The Lens	*source_urls*	4,142	82.9 (±1.4)	*is_open_access*	50,666	43.6 (±0.4)
Scilit	*pdf_url*	51,538	45.4 (±0.4)	*unpaywall_pdf_url*	28,841	25.4 (±0.3)
Semantic Scholar	*publicationVenue-url*	36,065	39.1 (±0.4)	*IsOpenaccess*	32,709	35.4 (±0.4)
OpenAlex		46,729	44.5 (±0.4)	*openaccess-oa_url*	49,190	42.6 (±0.4)

engines Google Scholar and Microsoft Academic stand out in this aspect because they only index documents that are accessible on the web. The remaining 19.2% of documents without links in Microsoft Academic is explained by the removal of handles. Färber and Ao (2022) detected more documents with links (94%), which would explain this difference, as well as the coverage of The Lens, because it also uses Microsoft Academic Graph as a source. In the case of Crossref, it could be because publishers deposit their landing pages to generate incoming traffic to their publications. Conversely, Semantic Scholar (39.1%) and Scilit (45.4%) provide fewer URLs, despite the fact that the former uses Crossref as a source. The reason is that Semantic Scholar only includes URLs of the venues, but not of the papers, and Scilit only indexes URLs with a pdf (*pdf_url*). This also occurs in Dimensions, where the proportion of publications with external links is the same as for open access articles (44.5%).

According to open access information, Google Scholar again identifies more open versions (52.2%), followed by Dimensions (44.5%), The Lens (43.6%), and OpenAlex (42.6%). These differences between Google Scholar and the other databases could be because Google Scholar indexes any open copy accessible on the web, regardless of whether the publications were released as open access or not (green open access). Conversely, Semantic Scholar (35.4%) and Scilit (25.4%) capture the fewest open documents. Figure 2 depicts two Venn

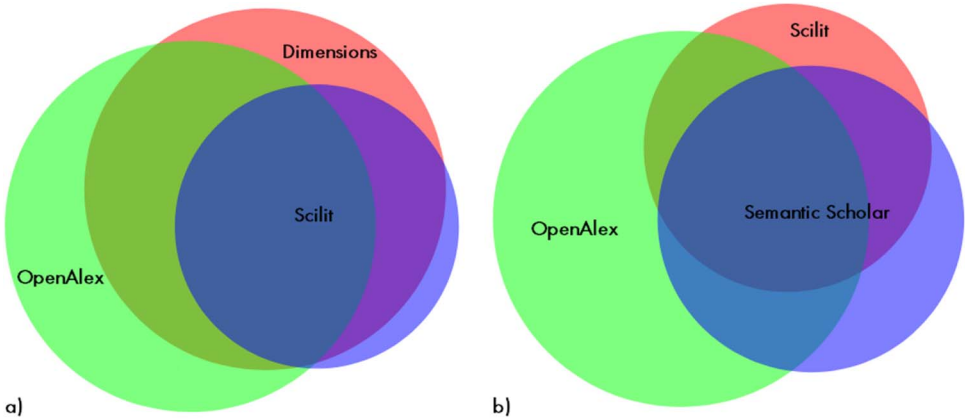

Figure 2. Overlap among databases identifying open access publications.

diagrams showing the overlap between databases according to open access records. Overall, the picture shows that, although the databases index a similar proportion of open access documents, the overlap is not significant. Figure 2(a) shows that OpenAlex and Dimensions share the largest proportion of records (81.1%), whereas OpenAlex and Scilit only have in common 46.1% of the records. This low overlap is surprising because both services use Unpaywall to detect open access publications. A possible explanation would be that Scilit only covers open access publications in pdf format (*unpaywall_pdf_url).* Semantic Scholar (Figure 2(b)) also shows a disparity with Scilit (49.8%) and OpenAlex (50%). In this case, Semantic Scholar uses its own criterion to detect open access publications.

5.3. Bibliographic Information

A critical element in a bibliographic database is the correct identification of indexed publications. Journal articles are identified using information that enables the document to be placed into the journal. Volume, issue, and pages are three fields that facilitate correct identification. All the databases include these fields, with the exception of Semantic Scholar, which has no field for issue.

Figure 3 and Table 5 depict the proportion of bibliographic data for journal articles in each database. Google Scholar is not included because it does not provide bibliographic information. In general, all the databases show high rates of completeness, including more information about volume than pages and issues. In this sense, Dimensions is again the platform with the highest completeness rates, with 100% of volume and 94.3% of pages, followed by The Lens with the highest number of pages (94.2%). The most noteworthy result is the low completeness degree of OpenAlex, with 50.2% of issue, 57.6% of pages, and 62.4% of volume. These figures are much lower than those reported by Microsoft Academic, its primary source. A manual inspection confirmed this lack of data, in which almost all the records ingested in December 2022 failed to include this information. This high completeness is also observable in the considerable overlap among confidence intervals, which shows that this information is essential in any scholarly database.

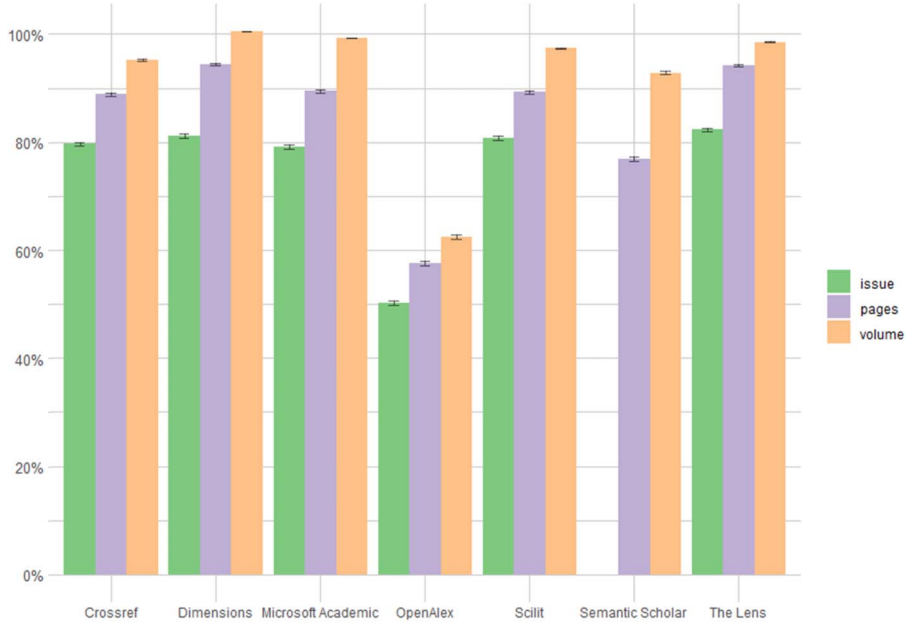

Figure 3. Proportion of bibliographic records with information about volume, pages, and issue.

Table 5. Percentage and number of articles with bibliographic info (volume, issue, and pages) in each database. Margin of error at 99%

Database	Publications	Volume	Volume %	Issue	Issue %	Pages	Pages %
Crossref	87,091	82,936	94.8 (±0.2)	69,406	79.3 (±0.4)	77,367	88.8 (±0.3)
Dimensions	83,760	83,760	100	68,011	81.2 (±0.3)	78,985	94.3 (±0.2)
Microsoft Academic	73,704	73,157	99.3 (±0.1)	58,301	79.1 (±0.4)	65,935	89.5 (±0.3)
The Lens	86,599	85,303	98.5 (±0.1)	71,297	82.3 (±0.3)	81,562	94.2 (±0.2)
Scilit	85,227	83,004	97.4 (±0.1)	68,801	80.7 (±0.3)	76,050	89.2 (±0.3)
Semantic Scholar	72,070	66,910	92.8 (±0.3)			55,367	76.8 (±0.4)
OpenAlex	87,081	54,333	62.4 (±0.4)	43,752	50.2 (±0.4)	50,191	57.6 (±0.4)

5.4. Document Type

Although journal articles make up more than 70% of the scientific literature, there is a large variety of scholarly documents (book, book chapters, conference papers, etc.) that also provide relevant scientific information and that are incorporated by many scholarly databases in their indexes. Scholarly databases categorize these typologies to inform about the academic nature of each item. However, the range of categories in each database varies significantly. For instance, although Crossref includes 33 document types, Dimensions summarizes its classification in only six classes (Table 6).

Table 6 displays the number of different document types and the number of records categorized in each database. Again, Google Scholar is excluded because the database has no document types. All the publications in Crossref (100%) and Dimensions (100%) are assigned to a typology, and OpenAlex (100%), The Lens (99.2%), and Scilit (99.8%) only find assignment problems in exceptional cases. This high completeness could be due to the fact that all these sources take document typologies from Crossref. Therefore, this lack of statistical differences could be more related to completeness than to data origin.

However, Microsoft Academic (80.3%) and Semantic Scholar (41.3%) present serious problems when classifying their records by typology. A possible explanation is that both search engines extract metadata from the web, and this information is not always available. Worth mentioning is the case of Semantic Scholar, which appears to use an automatic procedure

Table 6. Number of document typologies and completeness degree in each database. Margin of error at 99%

Database	Typologies	Publications	Percentage
Crossref	33	116,592	100
Dimensions	6	105,062	100
Microsoft Academic	7	77,389	80.3 (±0.3)
The Lens	17	115,396	99.2 (±0.1)
Scilit	20	113,168	99.8 (±0.04)
Semantic Scholar	12	38,096	41.3 (±0.4)
OpenAlex	13	115,853	99.98 (±0.01)

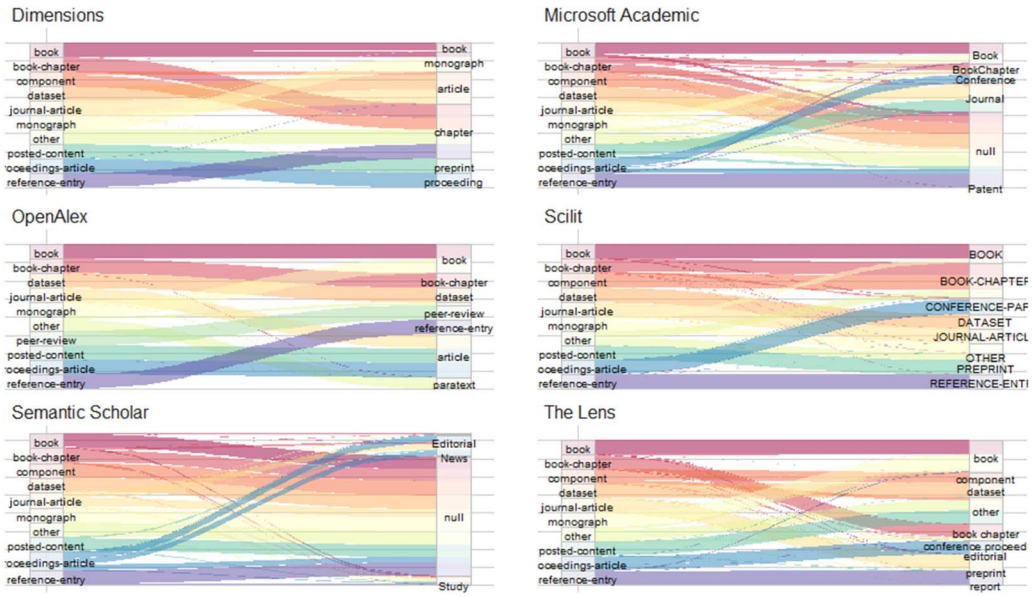

Figure 4. Alluvial graph with the transfer of document types between Crossref and the other databases. The stratum on the left shows the original Crossref classification and on the right the classification system of each database. Some labels were omitted to avoid overlaps.

to assign more than one typology based more on content criteria (*Review, Study, CaseReport*, etc.) than on formal criteria.

Figure 4 shows different alluvial graphs illustrating document type transfers between the Crossref classification and the systems of each database. The aim is to elucidate how each database assigns document types to their records. To improve the clarity of the graph, only the 10 most common categories in Crossref are displayed. For instance, Dimensions significantly reduces the document categories, integrating *book-chapter* (99.9%), *component* (78.7%), *reference-entry* (100%), and *other* (100%) into the *chapter* category, and *dataset* (100%) and *journal-article* (99.1%) into *article*. Microsoft Academic shows important problems when classifying *book chapters* (46.8%) and *proceeding-articles* (65.3%). Up to July 2023, OpenAlex directly used the Crossref scheme without any variation. Today, *monograph* (100%) is entered under *book*, and some marginal typologies (i.e., *others, component, journal, proceeding, journal-issue*) are grouped under the *paratext* category. Scilit also presents slight variations to the Crossref framework, creating *Conference-Paper* to group any document related to a conference, regardless of whether it is published as a *journal-article, book-chapter*, or *proceeding-article*. Semantic Scholar has serious problems when classifying most of the document typologies because only 46.2% of *proceeding-articles* are classified as *Conference* and 35.3% of *journal-articles* as *JournalArticle*. Finally, The Lens also shows similarities with the Crossref classification: *proceeding-articles* are split into *conference proceedings* (56.2%) and *conference proceeding articles* (35.7%), and *posted-content* is entered under *other* (94.4%).

5.5. Publication Dates

Electronic publishing has led to the appearance of multiple dates associated with the same document, describing different lifespan stages. This variety of dates also causes problems in the management of these publications (Ortega, 2022). Crossref includes more dates (up to eight), followed by Dimensions with five, and Microsoft Academic and The Lens with four. Crossref (*created*), Dimensions (*date_inserted*), Microsoft Academic (*CreatedDate*), and The

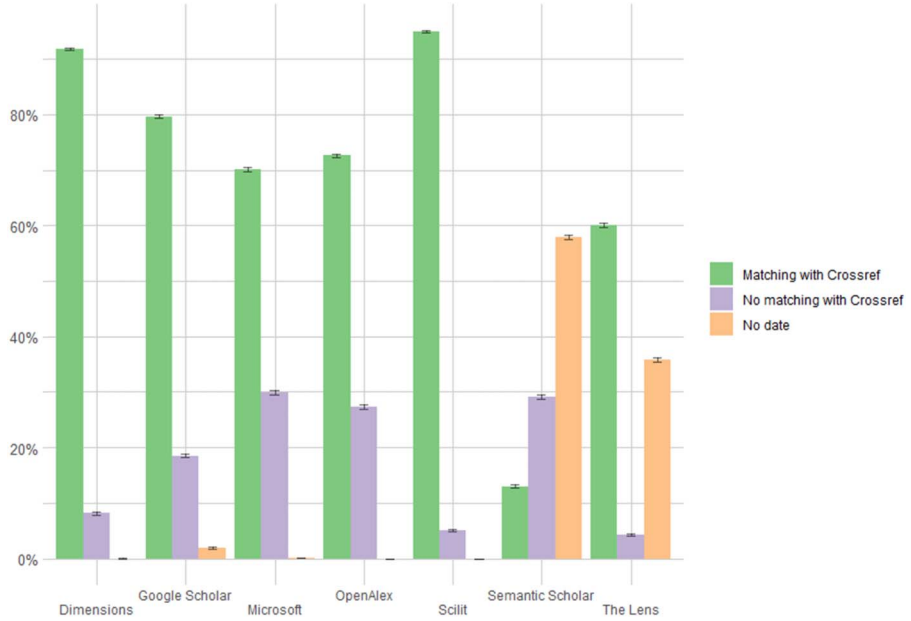

Figure 5. Percentage of publication dates matching with Crossref dates.

Lens (*created*) display the date when the record was created; and Crossref (*published-print, published-online*), Dimensions (*date_print, date_online*), and Scilit (*date_print, publication_year*) distinguish between the print and online date.

Publication date is common to all the databases and allows us to analyze the reliability of this information in each. A way to test the accuracy of these data is to compare the matching percentage with Crossref dates. Crossref is directly fed by publishers, which could make it the most authoritative source for exact and correct publication dates. Crossref fields that most match the publication date are *published* (88%), *published-online* (6%), and *created* (5%).

Figure 5 and Table 7 depict the proportion of publication dates that match some Crossref dates (i.e., *published-print, published-online, created, deposited, indexed,* and *issued*) and the percentage of publications with no date. Google Scholar only includes publication year. The comparison is thus done by year, not by date, resulting in a much higher match. That said, Google Scholar matches only 79.6%. The bar graph shows that Scilit (94.9%) and Dimensions (91.9%) have the best match with Crossref, and Microsoft Academic (70.1%) and OpenAlex (72.6%) have lower matching rates. These results could be explained because Dimensions and Scilit take their data from Crossref, and OpenAlex is an adaptation of the Microsoft Academic database. The most interesting result is perhaps the high proportion of publications with no date in Semantic Scholar (57.9%) and The Lens (35.8%). In the case of Semantic Scholar, it could be due to parsing problems when information is extracted from websites. For The Lens, however, this absence of information could be due to technical problems. It has the lowest proportion of no matching publication dates (4.3%), which could indicate that The Lens is also extracting the publication date from Crossref.

5.6. Language

A relevant factor to consider in a scholarly database is the language of the full text. The release of research documents in a language other than English is growing, and there is increasing demand for publications by local research communities. However, this information is only supplied by

Table 7. Percentage and number of articles with and without publication date matching and not matching with Crossref. Margin of error at 99%

Database	No date	No date %	No match with Crossref	No match with Crossref %	Match with Crossref	Match with Crossref %
Dimensions	2	0.0	8,512	8.1 (±0.2)	96,634	91.9 (±0.2)
The Lens	41,696	35.8 (±.36)	4,986	4.3 (±0.2)	70,076	60.0 (±0.4)
Microsoft Academic	95	0.1 (±.03)	29,064	29.8 (±0.4)	68,266	70.1 (±0.4)
Scilit	1	0.0	5,751	5.1 (±0.2)	107,668	94.9 (±0.2)
Semantic Scholar	53,434	57.9 (±.42)	26,892	29.1 (±0.4)	11,988	13.0 (±0.3)
OpenAlex	0	0.0	31,705	27.4 (±0.3)	84,178	72.6 (±0.3)
Google Scholar	1,900	1.9 (±.11)	18,809	18.5 (±0.3)	80,982	79.6 (±0.3)

Crossref (*language*), Microsoft Academic (*LanguageCode*), The Lens (*languages*), and Scilit (*language*). In our study, we extracted this information from Crossref, Scilit, and Microsoft Academic. Technical problems made it impossible to retrieve this information from The Lens. The results show that Scilit is the platform that identifies the language of most publications (99.9%), followed by Microsoft Academic (77.6%) and Crossref (57.1%). A manual inspection of the language assignment seems to indicate that Crossref assigns language according to venue, Scilit according to titles and abstracts, and Microsoft takes the language from webpage metadata.

Figure 6 displays a Venn diagram plotting the overlap between Crossref, Microsoft Academic, and Scilit when identifying the same publication language. The results show that Scilit largely matches Crossref (93.3%) and Microsoft Academic (89.8%), and the coincidence between Crossref (38%) and Microsoft Academic (54.9%) is low. These differences show how the methodological differences between Crossref (venues) and Microsoft Academic (webpages) influence language assignation, and how the use of content elements (title and abstract) improves language detection in Scilit.

5.7. Identifiers

Much of the current proliferation of scholarly databases is due to the consolidation of external identifiers that make it possible to individualize publications (duplicate management) and

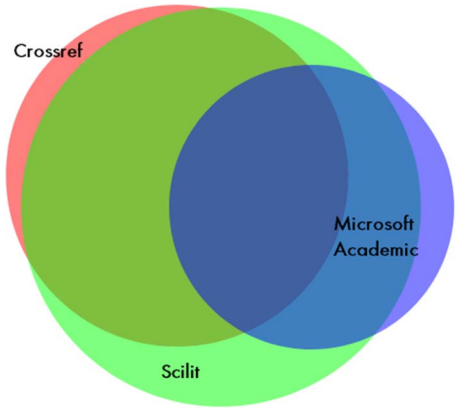

Figure 6. Overlap between databases assigning the same language.

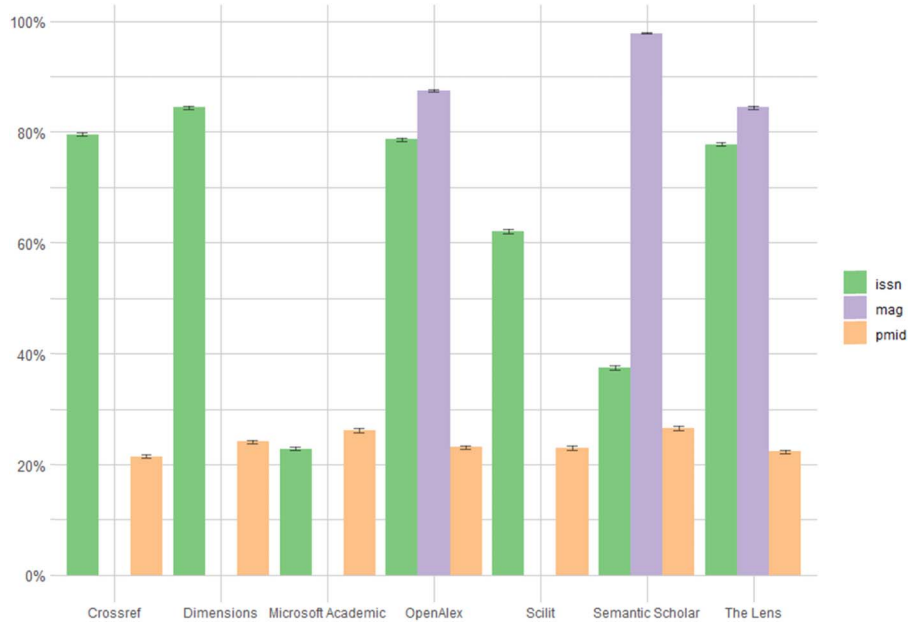

Figure 7. Percentage of different identifiers in each database.

connect with other sources, thereby enriching the information about publications. Apart from DOIs, many databases index different external identifiers. Semantic Scholar (*externalIds*), The Lens (*external_ids*), OpenAlex (*ids*), and Microsoft Academic (*AttributeType*) have a specific field for external identifiers. Crossref, Scilit, and Dimensions have different fields for each identifier. Google Scholar provides no identifier.

Figure 7 shows the proportion of the three most frequent identifiers (ISSN, MAG, and PMID) in each database. ISSN is the code that identifies journals and series, MAG is the Microsoft Academic identifier, and PMID is the PubMed ID. The aim is to discover the sources of these databases and to explore the identification of publication venues. The results show that all databases index or identify publications from PubMed to a similar extent, ranging from 21.5% for Crossref to 26.5% for Semantic Scholar (Table 8). Only three databases—OpenAlex, Semantic Scholar, and The Lens—take data from Microsoft Academic. Semantic Scholar (97.8%) and OpenAlex (87.4%) index the most publications. Their differences suggest, on the

Table 8. Percentage and number of articles with different identifiers. Margin of error at 99%

Database	PMID	PMID %	ISSN	ISSN %	MAG	MAG %
Crossref	25,026	21.5 (±0.3)	92,742	79.5 (±0.3)		
Dimensions	25,312	24.1 (±0.3)	88,673	84.4 (±0.3)		
Microsoft Academic	25,244	26.2 (±0.4)	22,043	22.9 (±0.4)		
The Lens	26,007	22.4 (±0.3)	90,467	77.8 (±0.3)	98,192	84.4 (±0.3)
Scilit	26,067	23 (±0.3)	70,418	62.1 (±0.4)		
Semantic Scholar	24,919	26.5 (±0.4)	34,563	37.4 (±0.4)	91,887	97.8 (±0.1)
OpenAlex	26,817	23.1 (±0.3)	91,104	78.6 (±0.3)	101,271	87.4 (±0.3)

one hand, that Semantic Scholar is highly dependent on Microsoft Academic and, on the other, that OpenAlex is already using other sources (Crossref) to expand its database. As for ISSN, all the services have a similar coverage of ISSNs (≈80%), although Microsoft Academic (22.9%) and Semantic Scholar (37.4%) have rather low proportions, suggesting deficient journal identification possibly deriving from the web extraction process.

6. DISCUSSION

This comparative analysis between free-access bibliographic databases has mainly focused on the completeness degree and quality of their metadata. This quantitative approach has made it possible to detect coverage limitations, allowing us to speculate about the technical reasons for these results. The study reports important outcomes about data sources and how information is processed. Overall, the results allow us to distinguish between academic search engines (Google Scholar, Microsoft Academic, Semantic Scholar) and third-party databases (Dimensions, The Lens, Scilit, OpenAlex). The former reveal clear problems with the number of fields that describe publications and their completeness. One explanation could be that these databases mainly obtain their data by crawling the web, and the information that webpages make available could be insufficient to correctly describe a publication. Thus, Semantic Scholar only includes abstracts for 54.5% of its publications and is the service with the lowest proportion of open access documents (35.4%) and external links (39.1%). Almost 60% of its publications have no document type, and it is the database that has the most publications with no publication date (57.9%). These figures reveal that Semantic Scholar has serious problems processing bibliographic information, which could be the root of the low quality of its metadata. In addition, the high proportion of records with the Microsoft Academic id (97.8%) and the similar coverage of PMIDs (Microsoft Academic = 26.2% (±0.4); Semantic Scholar = 26.5% (±0.4)) and no matching dates with Crossref (Microsoft Academic = 29.8% (±0.4); Semantic Scholar = 29.1% (±0.4)) suggests that the core of Semantic Scholar relies on Microsoft Academic and less on crawling the web.

To a lesser extent, Microsoft Academic also shows a low document type classification (79.5%), lack of information about open access publications, and the lowest proportion of ISSNs per publication (22.9%). According to document type, some recent studies observed that more than half of publications include this information (Färber & Ao, 2022; Visser et al., 2021). This disparity with our results could be because patents are not included in our study (approximately 20%). It does, however, highlight the coverage of external links (80%). These results are important for understanding how other products, based on their data (Semantic Scholar, The Lens, and OpenAlex), have inherited or solved these problems.

The main issue with Google Scholar is that it provides very little information about publications. Basic information, such as document type, bibliographic information, publication date, or identifiers, is missing from this database. With the exception of citations and versions, Google Scholar barely adds value to their records. However, as a search engine, it is the service that provides the most external links (97.1%) and detects the most open versions (52.2%), thereby confirming that it is the best gateway to scientific literature on the web (Martín-Martín et al., 2021).

Conversely, third-party databases supported on Crossref (Dimensions, The Lens, and Scilit) have more metadata details, with a high completeness degree. Dimensions could be considered the database that most enriches and improves information provided by Crossref. It indexes the most publication abstracts (69.6%), identifies the most open access articles (44.5%) (Basson et al., 2022), has the best coverage of bibliographic info (volume = 100%,

issue = 81%, pages = 92%), and 100% of publications are listed according to their typology. These results illustrate that Dimensions endeavors to improve Crossref metadata by adding abstracts, document typology, open access status, and so forth, to their records, with a high completeness rate. However, other sources, such as Scilit and The Lens, show signs of low data processing efficiency. For instance, Scilit is the commercial product that indexes the smallest number of abstracts (50.5%) and the lowest proportion of open access documents (25.4%). The Lens has reported serious problems with publication dates (35.8%). These findings reveal that the main risk of third-party databases is that they require a considerable processing effort to improve the quality of their metadata. A similar case is OpenAlex, which is based on both Microsoft Academic and Crossref. This integration of different sources would cause a loss of information, with a high proportion of missing bibliographic data (volume = 62%, issue = 50%, pages = 51%). The results for OpenAlex indicate that this new database is similar to Microsoft Academic (same proportion of abstracts and dates, the second database with the highest number of Microsoft Academic identifiers), but with the addition of DOIs and document types from Crossref (Scheidsteger & Haunschild, 2023). This active processing of different sources illustrates the importance of these tasks for offering a reliable scholarly bibliographic product (Priem, Piwowar, & Orr, 2022).

Any comparative analysis of bibliographic databases is always limited by the reference database used in the study. In our case, Crossref presents some particularities that should be considered in the interpretation of the results. Crossref is merely a repository of unprocessed publishers' data. Its coverage is determined by its partners, who decide on the data and how they are deposited, which could influence the representativeness of the sample and the quality of the metadata. A way of lessening this limitation is to make a random selection of the sample. This problem would also influence reliability because publishers may deposit incomplete records, lacking some fields and containing errors, including incorrect or null data about their own publications.

7. CONCLUSIONS

The results lead us to conclude that a random Crossref sample enables the comparison of a wide range of scholarly bibliographic databases, and the benchmarking of the amount of information and completeness degree of these databases, with regard to different facets. Accordingly, the numbers of publications with abstract, external links, open access status, document type, or publication date have been measured across databases. Extraction of the same data from each database has shed light on overlaps, which has led us to identify possible connections between databases.

The results show that databases based on external sources can generate more and improved metadata than academic search engines extracting information from the web. Search engines have the power to reach distant publications and detect more open copies, but they lack the ability to retrieve reliable descriptive data about publications from webpages. However, this integration of different sources also produces problems, such as the loss of information (The Lens and publication date, or OpenAlex and bibliographic info), or limitations inherited from the primary sources (OpenAlex and the publication dates of Microsoft Academic).

Finally, Dimensions provides the greatest number of fields about publications and the highest completeness degree. OpenAlex, The Lens, and Scilit also include a varied range of fields but display certain integration problems, with a lack of information and low completeness rates in specific fields. Conversely, search engines such as Semantic Scholar and Google Scholar lack important fields for identifying and searching publications (document types,

certain bibliographic info). Microsoft Academic is the search engine that provides the most publication fields, and its completeness rate is high, though it lacks information on open access status and document type for some publications.

AUTHOR CONTRIBUTIONS

Lorena Delgado-Quirós: Data curation, Formal analysis, Resources, Software. José Luis Ortega: Conceptualization, Funding acquisition, Investigation, Methodology, Writing—original draft, Writing—review & editing.

COMPETING INTERESTS

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

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.

DATA AVAILABILITY

For legal reasons, data from many of the databases (Dimensions, Google Scholar, Scilit, and The Lens) cannot be made openly available. Data from Crossref, OpenAlex, Microsoft Academic, and Semantic Scholar are openly available because they have been released under a CC-BY license. We have uploaded the instructions on how to retrieve the data in each database (https://osf.io/yw6j4). In this form, readers with credentials can download the data and reproduce the study.

REFERENCES

Basson, I., Simard, M. A., Ouangré, Z. A., Sugimoto, C. R., & Larivière, V. (2022). The effect of data sources on the measurement of open access: A comparison of Dimensions and the Web of Science. *PLOS ONE*, *17*(3), e0265545. https://doi.org/10.1371/journal.pone.0265545, PubMed: 35358227

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/

Bruce, T., & Hillman, D. (2004). The continuum of metadata quality: Defining, expressing, exploiting. In D. Hillman & E. Westbrooks (Eds.), *Metadata in practice* (pp. 238–256). Chicago, IL: ALA Editions.

Färber, M., & Ao, L. (2022). The Microsoft Academic Knowledge Graph enhanced: Author name disambiguation, publication classification, and embeddings. *Quantitative Science Studies*, *3*(1), 51–98. https://doi.org/10.1162/qss_a_00183

Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2016). Empirical analysis and classification of database errors in Scopus and Web of Science. *Journal of Informetrics*, *10*(4), 933–953. https://doi.org/10.1016/j.joi.2016.07.003

Google Scholar. (2023). *Inclusion guidelines for webmasters*. https://scholar.google.com/intl/es/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*, 593494. https://doi.org/10.3389/frma.2020.593494, PubMed: 33870055

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

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. https://doi.org/10.1162/qss_a_00022

Herrmannova, D., & Knoth, P. (2016). An analysis of the Microsoft Academic Graph. *D-Lib Magazine*, *22*(9/10), 37. https://doi.org/10.1045/september2016-herrmannova

Herzog, C., Hook, D., & Konkiel, S. (2020). Dimensions: Bringing down barriers between scientometricians and data. *Quantitative Science Studies*, *1*(1), 387–395. https://doi.org/10.1162/qss_a_00020

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

Metrics and Analytics, *3*, 23. https://doi.org/10.3389/frma.2018.00023

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

Jefferson, O. A., Koellhofer, D., Warren, B., & Jefferson, R. (2019). *The Lens MetaRecord and LensID: An open identifier system for aggregated metadata and versioning of knowledge artefacts*. https://osf.io/preprints/lissa/t56yh/

Kramer, B., & de Jonge, H. (2022). The availability and completeness of open funder metadata: Case study for publications funded by the Dutch Research Council. *Quantitative Science Studies*, *3*(3), 583–599. https://doi.org/10.1162/qss_a_00210

Kubler, S., Robert, J., Neumaier, S., Umbrich, J., & Le Traon, Y. (2018). Comparison of metadata quality in open data portals using the Analytic Hierarchy Process. *Government Information Quarterly*, *35*(1), 13–29. https://doi.org/10.1016/j.giq.2017.11.003

Liu, W., Hu, G., & Tang, L. (2018). Missing author address information in Web of Science—An explorative study. *Journal of Informetrics*, *12*(3), 985–997. https://doi.org/10.1016/j.joi.2018.07.008

Lutai, A. V., & Lyubushko, E. E. (2022). *Comparison of metadata quality in CrossRef, Lens, OpenAlex, Scopus, Semantic Scholar, Web of Science Core Collection databases*. Russian Foundation for Basic Research. https://podpiska.rfbr.ru/storage/reports2021/2022_meta_quality.html

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. https://doi.org/10.1007/s11192-018-2820-9

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, PubMed: 32981987

Ochoa, X., & Duval, E. (2009). Automatic evaluation of metadata quality in digital repositories. *International Journal on Digital Libraries*, *10*, 67–91. https://doi.org/10.1007/s00799-009-0054-4

Ortega, J. L. (2022). When is a paper published? *The Research Whisperer*. https://researchwhisperer.org/2022/02/08/when-is-a-paper-published/

Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv*. https://doi.org/10.48550/arXiv.2205.01833

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

Ranjbar-Sahraei, B., & van Eck, N. J. (2018). Accuracy of affiliation information in Microsoft Academic: Implications for institutional level research evaluation. In *STI 2018 Conference Proceedings* (pp. 1065–1067). Centre for Science and Technology Studies (CWTS).

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). https://doi.org/10.3145/epi.2023.mar.09

Tani, A., Candela, L., & Castelli, D. (2013). Dealing with metadata quality: The legacy of digital library efforts. *Information Processing & Management*, *49*(6), 1194–1205. https://doi.org/10.1016/j.ipm.2013.05.003

Valderrama-Zurián, J. C., Aguilar-Moya, R., Melero-Fuentes, D., & Aleixandre-Benavent, R. (2015). A systematic analysis of duplicate records in Scopus. *Journal of Informetrics*, *9*(3), 570–576. https://doi.org/10.1016/j.joi.2015.05.002

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* [online]. 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

Wade, A. D. (2022). The Semantic Scholar Academic Graph (S2AG). In *Companion Proceedings of the Web Conference 2022* (p. 739). https://doi.org/10.1145/3487553.3527147

Waltman, L., Kramer, B., Hendricks, G., & Vickery, B. (2020). Open abstracts: Where are we? *Crossref Blog*. https://www.crossref.org/blog/open-abstracts-where-are-we/

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