Quantitative comparison of publication metadata in eight free-access databases

Lorena Delgado-Quirós (ORCID: 0000-0001-8738-7276) and José Luis Ortega[1] (ORCID: 0000-0001-9857-1511)

Institute for Advanced Social Studies (IESA-CSIC), Córdoba, Spain

Joint Research Unit Knowledge Transfer and Innovation, (UCO-CSIC), Córdoba, Spain

jortega@iesa.csic.es ; ldelgado@iesa.csic.es

Abstract

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

Keywords: academic search engines, metadata quality, third-party databases, scholarly bibliographic databases, open access

1. Introduction

The recent proliferation of bibliographic scholarly databases has stimulated the interest in these new platforms and their possibilities to find scientific literature and provide different bibliometric indicators. This attention has been focused on testing the performance of these new systems in relation to traditional products such as citation indexes (i.e. Web of Science, Scopus) and academic search engines (i.e. Google Scholar, Microsoft Academic). These new products could be defined as hybrid databases because they share characteristics with the former ones. 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 by a free access model in which users do not require

[1] Corresponding author: Institute for Advanced Social Studies (IESA-CSIC), Camposanto de los Mártires, 7 14004 Córdoba, Spain jortega@iesa.csic.es

subscription fee to search and retrieve documents. Even in some cases, they provide open data through Rest APIs or dump files.

However, these hybrid products have some particularities that make them different. The most important is that they are fed by third party sources. The appearance of Crossref as 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, DOAJ, repositories) have made possible the emergence of these bibliographic products that, quickly and with a low cost, coverage large part 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 that avoids the appearance of duplicated record, filters non-scholarly materials, and manages different versions of the same document. This also influences the quality of their metadata because they are the result of the integration of external and internal descriptions.

Due to this, the study about the quality of the publication metadata in the new scholarly databases allows us to appreciate to what extent these processing efforts are accomplished and to value the suitability and reliability of these search tools for 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 (Van Eck et al., 2018; Gusenbauer, 2019; Martín-Martín et al., 2021; Visser et al., 2021). This quantitative procedure is excessively centered on the size of each platform and leaves aside 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 to inform about the richness and limitations of those sources. Hendricks et al. (2020) described the working of Crossref database and analyzed the completeness of their metadata. Similar papers were published describing Semantic Scholar (Wade, 2022), Lens (Jefferson et al., 2019), Dimensions (Herzog et al., 2020) and Microsoft Academic (Wang et al., 2020). Many of these studies were descriptive review written by their employeers without a critical discussion about the quality of their data.

In other cases, the coverage of certain elements or entities in different scholarly databases have been studied to test their performance in processing specific information. Hug and Brändle (2017) analyzed in detail the coverage of Microsoft Academic, and they found important problem in the assignation 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 et al. (2021) compared affiliation information between Scopus and Dimensions, and they 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. He found that as larger is a database more disambiguation problems show. Kramer and de Jonge

(2022) analyzed the information about funders included by Crossref, Lens, WoS, Scopus and Dimensions, finding important differences when they come 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 indexations of Russian authors.

Regarding to publications, some studies have explored the information amount and quality of this key entity in scholarly databases. Herrmannova and Knoth (2016) tested the reliability of the publication date in the Microsoft Academic Graph and they found that 88% of cases showed a correct date. Liu et al. (2018) detected that approximately 20% of WoS publications have no information from the address field. Basson et al. (2022) showed that the proportion of open access documents in Dimensions is higher than WoS because the first one indexes more publications from Asian and Latin-American countries. Other studies have revised errors and inconsistences in different academic databases to test their suitability for bibliometric studies or just for bibliographic searches. Thus, some articles have analyzed duplicated records management in Scopus (Valderrama-Zurián et al., 2015) and WoS (Franceschini et al., 2016).

3. Objectives

The main objective of this study is to compare the metadata quality about research publications in the new academic databases using a quantitative approach, with the aim of describing the advantages and limitations of these scholarly platforms providing bibliographic information for analytical studies and secondary products. To this end, seven of these new bibliographic databases were analyzed, considering 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? And therefore, to value the information richness of these databases?
- Do similarities or discrepancies among databases allows us to delimit different models of databases, with their advantages and limitations?
- Which do databases provide the most metadata and they have a higher completeness rate?

4. Methods

4.1. Source selection criteria

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

- They have to be freely accessible through the Web: it means a free-subscription search interface.
- They also provide metrics for research evaluation.

4.2. Sample selection and extraction

Crossref was selected as control sample due to several causes. The first one is due to an operational question. Crossref is a publishers' consortium that assigns the Document Object Identifier (DOI), the most extended persistent identifier of research publications in the publishing system. Although their coverage is limited to only publisher members (Visser et al., 2019), its use is justified because all these platforms allow to query publications by DOIs, favoring a rapid and exact matching. The second reason is related to methodological issues, Crossref is the only service that provides the extraction of random samples of documents (https://api.crossref.org/works?sample=100). This fact reinforces the representativeness of the sample, 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 to any published material, regardless of typology, discipline or language. This means that Crossref database does not have any inclusion criteria that could limit the coverage of certain types of documents (e.g. indexes, acknowledgements, front covers). This non-selective criterion would lead us to clearly appreciate the inclusion policies of the different bibliographic platforms. Finally, Crossref is fed by publishers, which deposit metadata about their publications. They could be considered the most authoritative source about the reliability and accuracy of their own publications.

4.3. Data retrieving

A sample of 116,648 DOIs were randomly extracted from Crossref in August 2020 and July 2021 with the only limitation of documents published between 2014 and 2018. This time window was selected in order to publications can accrue a significant number of citations and other metrics. This sample was generated performing 1,200 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 resulting distribution by document type coincides with the entire database (Hendricks et al., 2020), which reinforce the reliability of the sample.

Next, this control sample was queried to each platform to match the records and extract all the information about each publication. This task was carried through July 2021, excepting 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 due to its novelty as open bibliographic source. The extraction process in each platform is described in detail:

- **Dimensions**: This database was accessed through their API (https://app.dimensions.ai/dsl/v2). A R package (i.e. dimensionsR[2]) 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.
- **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[3] R package was used to emulate a browser session and avoid anti-robot actions (i.e.

[2] https://github.com/massimoaria/dimensionsR

[3] https://docs.ropensci.org/RSelenium/

captchas). As it is possible that some DOIs could be not indexed (Martín-Martín, 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**: Several methods were used to obtain the coverage of this service. 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[4], a R package, was used to query the API. However, the low indexation of DOIs (37.1%) and that these were case sensitive, made us to download the entire table of publications available in Zenodo (https://zenodo.org/record/2628216) and locally match with the sample, using DOIs and titles.
- **Scilit**: this platform was accessed using a public API (https://app.scilit.net/api/v1/). Because the access must be done using 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[5] R package was used to extract the data. However, API was directly queried after to detect some problems in the retrieval process. A python script was written.
- **The 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. 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 decided then to extract a little sample of 5,000 records directly from the main search page (https://www.lens.org/lens/) to supply this limitation in January 2023.

This study has a qualitative-quantitative approach, in which we extract large data samples from different sources to then compare the quantity and quality of the included information. API documentation about each database was analyzed to know data available about publications.

Table 1. web source with information about publication metadata in each database

Database	**Pub.**	**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
Lens	116,337 (4,996)	https://docs.api.lens.org/response-scholar.html
OpenAlex	115,881	https://docs.openalex.org/

[4] https://docs.ropensci.org/microdemic/

[5] https://github.com/njahn82/semscholar

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 to the description of research publications indexed in these databases. Publications are the central element in the publishing ecosystem and they 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. Due to this, publications are the entity with more fields, going from the 38 fields in Crossref to the 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 the publication content because it provides a summary of the research. All the analyzed databases index this element. In the case of Microsoft Academic, the table with this information (*PaperAbstractsInvertedIndex*) is not already available. Although, early studies detected a coverage of 58% (Färber & Ao, 2022). Google Scholar does not exactly index the abstract of the publication, but it extracts parts of the document text (Google Scholar, 2023).

Table 2. Proportion of publications with abstract in each database

Database	**field**	**pub.**	**pub. %**
Crossref	*abstract*	15,927	13.66%
Dimensions	*abstract*	73,145	69.62%
Google Scholar		73,899	91.66%
The Lens	*abstract*	3,133	62.7%
Scilit	*abstract*	57,300	50.52%
Semantic Scholar	*abstract*	50,263	54.45%
OpenAlex	*abstract_inverted_index*	73,899	63.77%

Table 2 shows the number and proportion of publications with abstract. Google Scholar is the database that indexes more articles with a summary (91.66%), although some of them are just an extraction of the text. Taking apart this, Dimensions is the database that indexes more articles with its pertinent abstract (69.6%), followed by OpenAlex (63.8%) with a similar proportion. Contrarily, Crossref is the database with less publications with abstract (13.7%). This last percentage is a little bit lower than the reported by Waltman et al. (2020) (21%), due, perhaps, to that our study also gathers other materials such book chapters and conferences 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 and the indexation services need to process documents to obtain this data. This fact would explain the overall low availability of abstracts in free-access databases, highlighting the cases of Scilit (50.5%) and Semantic Scholar (54.45%).

5.2. *Access*

Today, a positive feature of scholarly databases is that they provide some type of access to original publications. The widespread electronic publishing allows to provide links to different venues where the document, partially or fully, is hosted. All the databases include external links to the original publication. Microsoft Academic and Crossref do not have a specific field for open access publications. Perhaps, the most problematic database is OpenAlex because it includes up to three fields (*landing_page_url*, *pdf_url*, and *oa_url*) with links to the original publication. An analysis of the content of those fields disclosed that *landing_page_url* in fact only includes DOI links, while *pdf_url* include similar information than *oa_url*. Then, we have considered that OpenAlex includes external links for only open access publications (*oa_url*). This also happen with Dimensions, which only indexes external links (*linkout*) for open_access (*open_access*) articles.

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

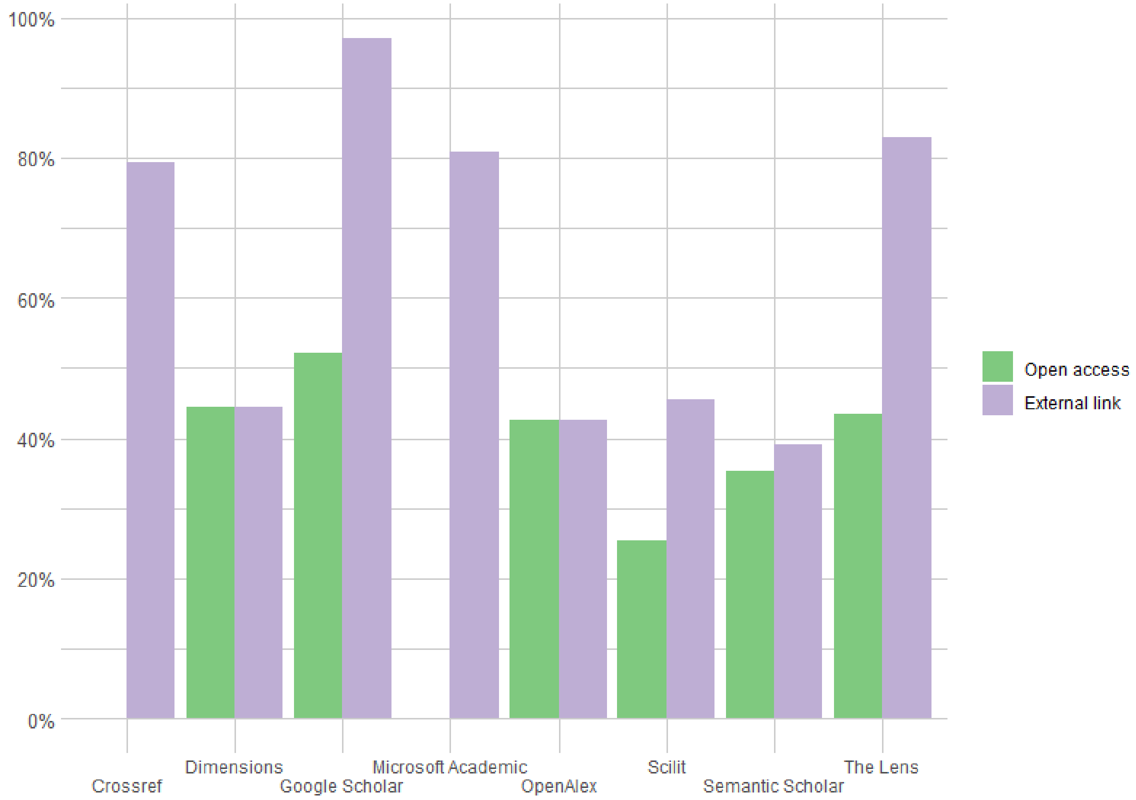


Table 3. Fields, publications and percentage of publications with external links and information about open access by database

	External links			Open access		
Database	**Field**	**Pub.**	**Pub. %**	**Field**	**Pub.**	**Pub. %**
Crossref	*Link*	92,561	79%			
Dimensions	*Linkout*	46,732	44.48%	*open_access*	46,729	44.48%
Google Scholar		98,714	97.1%		53,034	52.2%
Microsoft Academic	*PaperURL*	77,877	80.8%			
The Lens	*source_urls*	4,142	82.9%	*is_open_access*	50,666	43.55%
Scilit	*pdf_url*	51,538	45.4%	*unpaywall_pdf_url*	28,841	25.43%

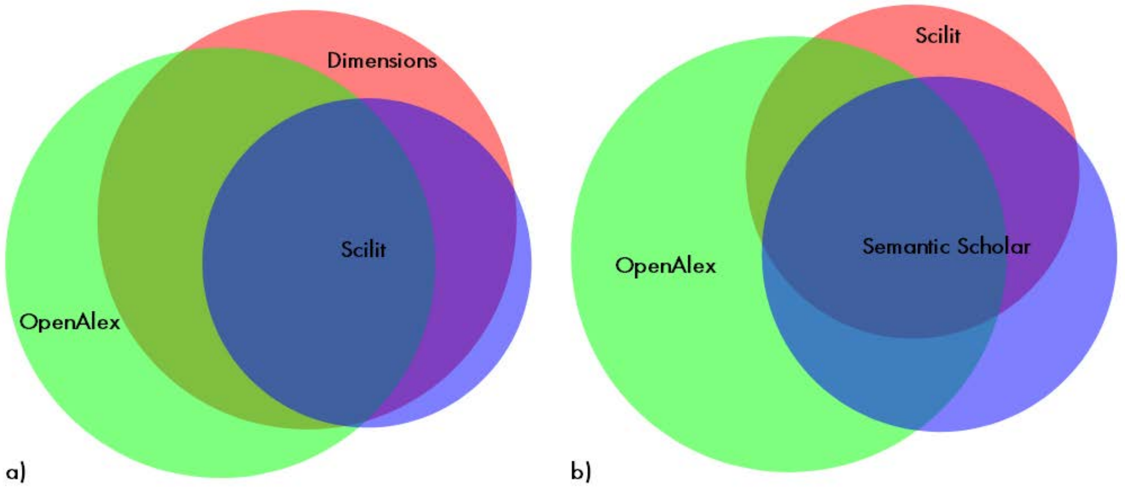
Semantic Scholar	*publicationVenue-url*	36,065	39.1%	*IsOpenaccess*	32,709	35.4%
OpenAlex		46,729	44.48%	*openaccess-oa_url*	49,190	42.6%

Figure 1 and Table 3 depicts the percentage of publications with external links to the original source and information about if they are open access or not by database. Google Scholar (97.1%) is the database that includes the most external links (97.1%), followed by The Lens (82.9%), Microsoft Academic (80.8%) and Crossref (79%). It is evident that academic search engines, Google Scholar and Microsoft Academic, highlight in this facet 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 removing of handles. Färber and Ao (2022) detected more documents with link (94%), which would explain this difference. This also would explain the coverage of The Lens, because it also uses Microsoft Academic Graph as source. In the case of Crossref could be due to publishers deposit their landing pages to generate incoming traffic to their publications. Contrarily, Semantic Scholar (39.1%) and Scilit (45.4%) provide less urls, in spite of the former one uses Crossref as source. The reason is that Semantic Scholar only include urls of the venues, but not of the papers; and Scilit only indexes urls with pdf (*pdf_url*). This same occurs in Dimensions where the proportion of publications with external links is the same than 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 due to Google Scholar indexes any open copy accessible on the Web, regardless of the publications were released as open access or not (green open access). In the contrary side, Semantic Scholar (35.4%) and Scilit (25.4%) capture the fewest open documents.

Figure 2 depicts two Venn 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 very high. 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. Semantic Scholar (Figure 2.b) also shows disparity with Scilit (49.8%) and OpenALex (50%).

Figure 2. Overlap among databases identifying open access publications



5.3. *Bibliographic info*

A critical element in a bibliographic database is the correct identification of the indexed publications. In the case of journal articles, this identification is done using information that allows us to place the document into the journal. Volume, issue and pages are three fields that make possible a correct identification. All the databases include these fields, excepting Semantic Scholar that does not have a field for issue.

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

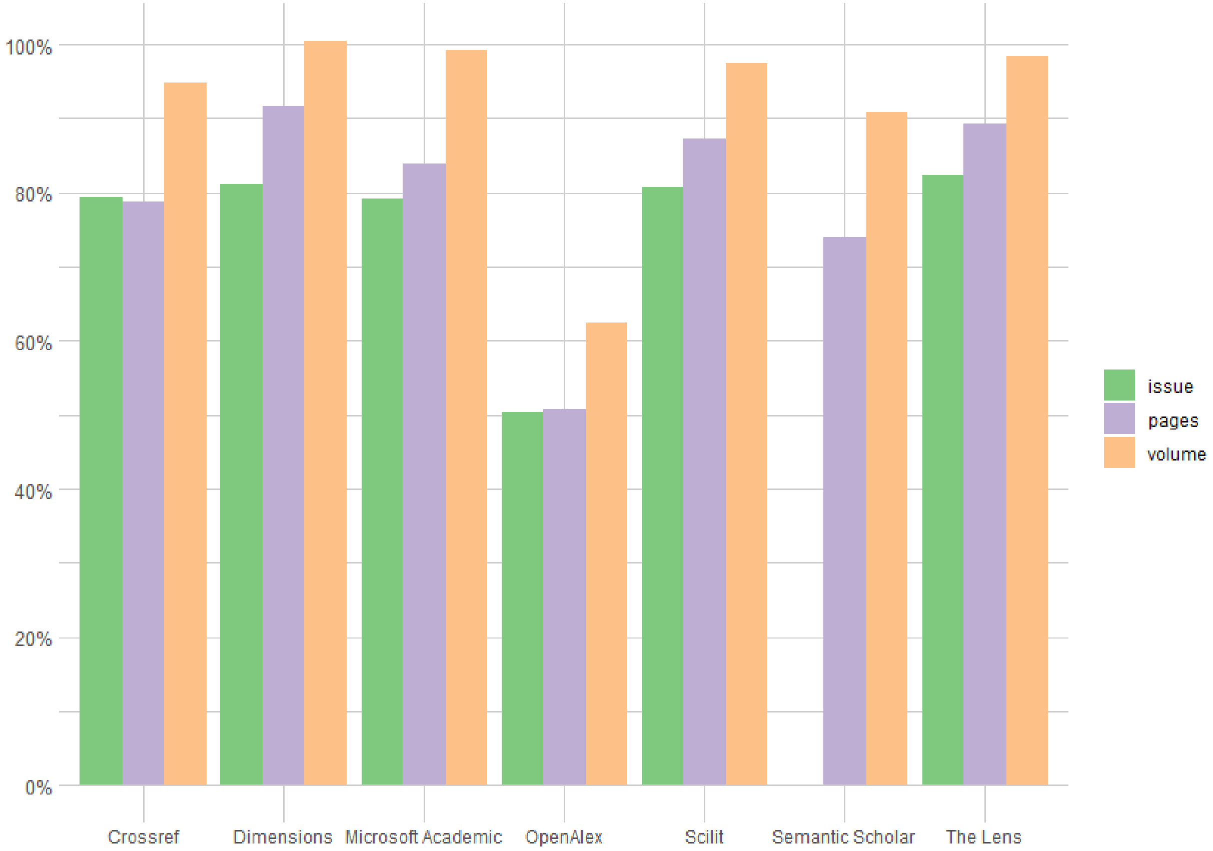


Table 4. Percentage and number of articles with bibliographic info (volume, issue and pages) in each database

Database	**Pub.**	**Volume**	**Volume %**	**Issue**	**Issue %**	**Pages**	**Pages %**
Crossref	87,524	82,936	94.8%	69,406	79.3%	91,889	78.8%
Dimensions	83,760	83,760	100%	68,011	81.2%	96,411	91.8%
Microsoft Academic	73,704	73,157	99.3%	58,301	79.1%	80,807	83.9%
The Lens	86,599	85,303	98.5%	71,297	82.3%	103,954	89.4%
Scilit	85,227	83,004	97.4%	68,801	80.7%	98,917	87.2%
Semantic Scholar	72,070	66,910	92.8%		0.0%	70,049	97.2%
OpenAlex	87,081	54,333	62.4%	43,752	50.2%	58,688	50.6%

Figure 3 and Table 4 depicts 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 that has highest completeness rates with 100% of volume and 91.8% of pages, followed by The Lens with the highest number of pages (82.3%). The most noteworthy result is the low completeness degree of OpenAlex, with 50.2% of issue, 50.6% of pages and 62.4% of volume. These figures are much lower than the 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 did not include this information.

5.4. *Document type*

Although more than 70% of the scientific literature are journal articles, there is a large variety of scholarly documents (book, book chapters, conference papers, etc.) that also provide relevant scientific information, and that many scholarly databases incorporate to 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, while Crossref includes 33 document types, Dimensions summarizes its classification to only six classes (Table 5).

Table 5. Number of document typologies and completeness degree in each database

Database	**Typologies**	**Pub.**	**Pub. %**
Crossref	33	116,592	100%
Dimensions	6	105,062	100%
Microsoft Academic	7	77,389	79.5%
The Lens	17	115,396	99.85%
Scilit	20	113,168	99.78%
Semantic Scholar	12	38,096	41.69%
OpenAlex	33	115,853	99.98%

Table 2 displays the number of different document types and the number of records categorized in each database. Again, Google Scholar is excluded because this database does not have document types. All the publications in Crossref (100%) and Dimensions (100%) are assigned to a typology, and OpenAlex (100%), The Lens (99.9%) and Scilit (99.8%) only find assignation problems in exceptional cases. However, Microsoft Academic (79.5%) and Semantic Scholar (41.7%) present serious problems to classify their records by typology. A possible explanation is that both search engines extract metadata from the Web, and this information is not always available. It is worth to mention the case of Semantic Scholar that seems that use an automatic procedure to assign more than one typology based more on content criteria (*Review*, *Study*, *CaseReport*, etc.) rather than on formal ones.

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

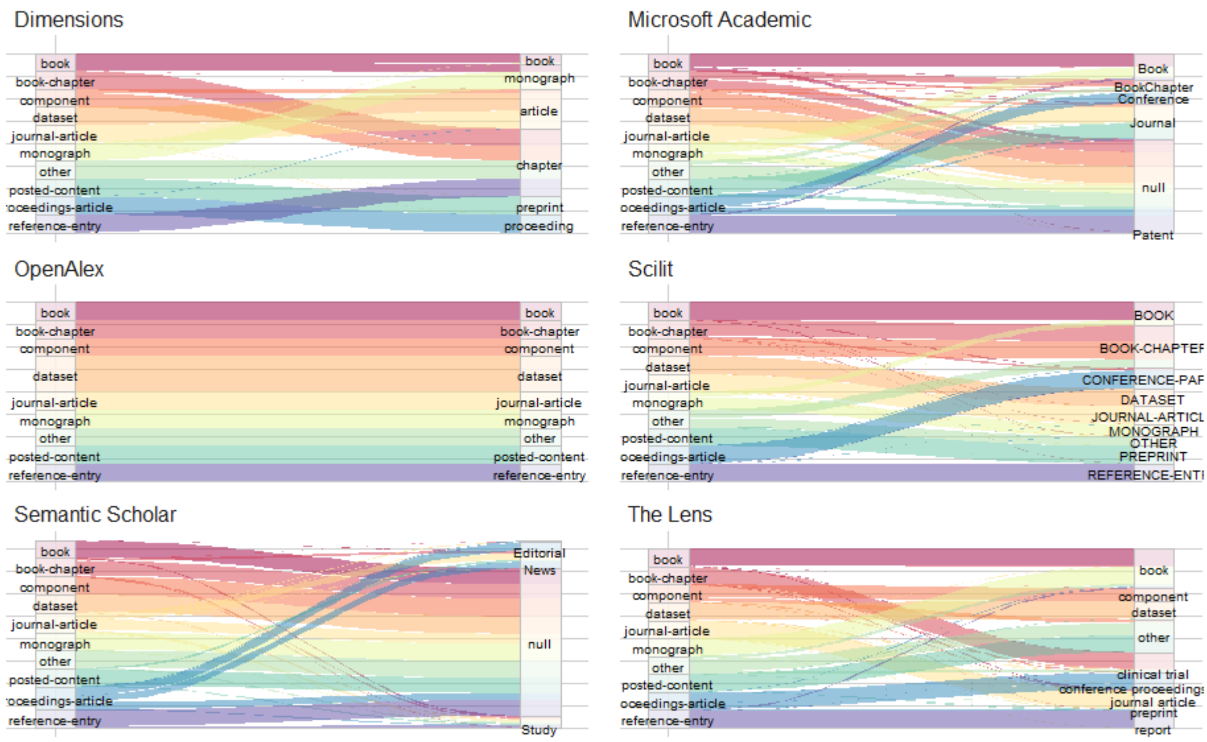


Figure 4 shows different alluvial graphs illustrating the document types transfers between Crossref classification and the systems of each database. The aim is to elucidate how each database assign document types to their records. To improve the clarity of the graph, only the ten most frequent categories in Crossref were displayed. For instance, Dimensions reduces significantly the document categories, integrating *book-chapter* (99.9%), *component* (78.7%), *reference-entry* (100%) and *other* (100%) into *chapter* category, and *dataset* (100%) and *journal-article* (99.1%) into *article*. Microsoft Academic shows important problems to classify *book chapters* (46.8%) and *proceeding-articles* (65.3%). OpenAlex directly uses the Crossref's scheme without any variation, while Scilit also presents slight variations to the Crossref's framework. Semantic Scholar has serious problems to classify 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's classification, and we can only highlight that *proceeding-articles* are split in *conference proceedings* (56.2%) and *conference proceeding articles* (35.7%), and *posted-content* is integrated in *other* (94.4%).

5.5. *Publication dates*

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

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

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

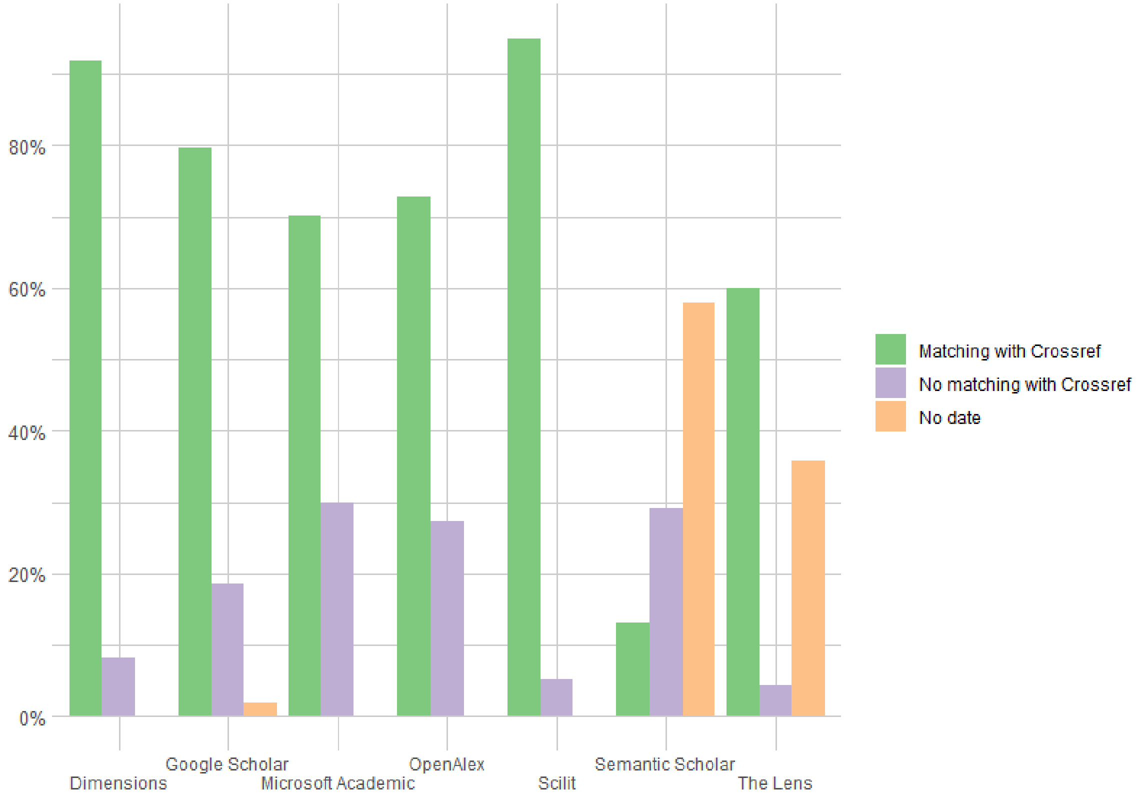


Table 6. Percentage and number of articles with and without publication date that match and no match with Crossref

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

Figure 5 and Table 6 depicts the proportion of publication dates that match with some of the Crossref's dates (i.e., *published-print*, *published-online*, *created*, *deposited*, *indexed* and *issued*) and the percentage of publications without date. Google Scholar only includes publication year, then the comparison is done with year, not with date. This cause that the agreement is much higher. Even so, Google Scholar just matches 79.6%. The bar graph shows that Scilit (94.9%) and Dimensions (91.9%) have the best matching with Crossref, while 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, while OpenAlex is an adaptation of the Microsoft Academic database. Perhaps, the most interesting result is the high proportion of publications without date in Semantic Scholar (57.9%) and The Lens (35.8%). In the case of Semantic Scholar could be due to parsing problems extracting information from web sites. Whereas, in the case of The Lens, this absence of information could be due to technical problems because it has the lowest proportion of no matching publication dates (4.3%), which could evidence 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, due to the growing releasing of research documents in a language distinct from the English one, and the increasing demand of publications by local research communities. However, this information is only supplied by Crossref (*language*), Microsoft Academic (*LanguageCode*), The Lens (*languages*) and Scilit (*language*). In our study, we extracted this information from Crossref, Scilit and Microsoft Academic, being impossible taking this information from The Lens due to technical problems. The results show that Scilit is the platform that identifies the language of the most publications, with a 99.9%. Followed by Microsoft Academic (77.6%) and Crossref (57.1%). A manual inspection of the language assignation seems indicate that Crossref assigns language according to the venues, Scilit according to titles and abstracts and Microsoft takes the language from the webpage metadata.

Figure 6. Overlap between databases assigning the same language

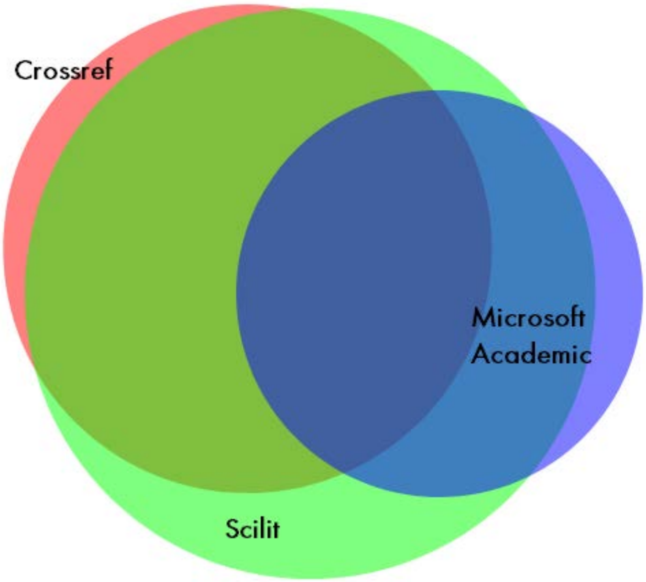


Figure 6 displays a Venn diagram plotting the overlap between Crossref, Microsoft Academic and Scilit identifying the same publication language. The results show that Scilit, to a large extent, matches with Crossref (93.3%) and Microsoft Academic (89.8%), while Crossref (38%) and Microsoft Academic (54.9%)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

5.7. Identifiers

A large part of the current proliferation of scholarly databases is due to the consolidation of external identifiers that make possible to individualize publications (duplicate management) and connect with other sources, 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 by each identifier. Google Scholar does not provide any identifier.

Figure 7. Percentage of different identifiers in each database

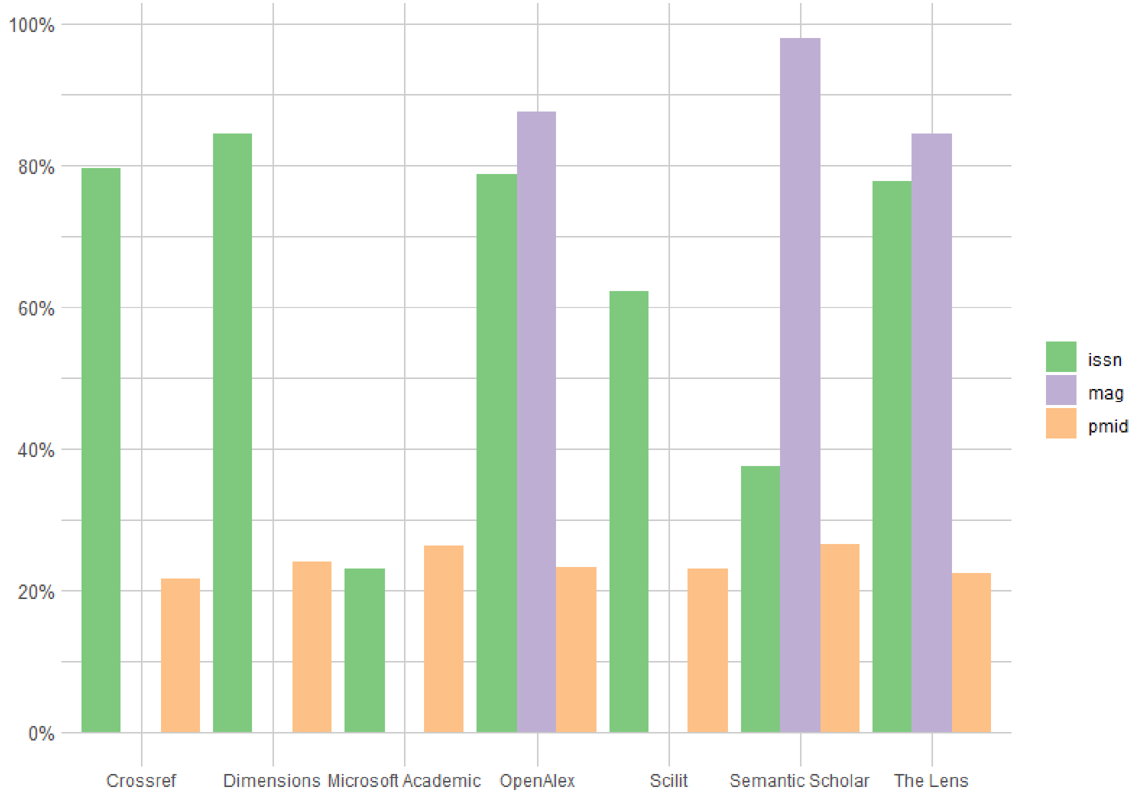


Figure 7 shows the proportion of the three most frequent identifiers (i.e. ISSN, MAG and PMID) in each database. ISSN is the code to identify journals and series, MAG is the Microsoft Academic's identifier and PMID is the PubMed ID. The aim is to know which are the sources of these databases and to explore the identification of publication venues. The results show that all databases index or identify publications from PubMed with the similar proportion, going from the 21.5% of Crossref to the 26.5% of Semantic Scholar. Only three databases, OpenAlex, Semantic Scholar and The Lens, take data from Microsoft Academic, being Semantic Scholar (97.8%) and OpenAlex (87.4%) the databases that index the most publications. Their differences suggest, on the one hand, that Semantic Scholar is a product highly dependent on Microsoft Academic and, on the other hand, OpenAlex is already using other sources (Crossref) to ingest its database. Regarding to ISSN, all the services has a similar coverage of ISSNs (≈80%), highlighting Microsoft Academic (22.9%) and Semantic Scholar (37.4%) with rather low proportions, which suggests a deficient journals identification possibly derived from the web extraction process.

6. Discussion

This comparative analysis between free-access bibliographic databases has reported important results about what are their data sources and how they process the information. Overall, the results allow to distinguish between academic search engines (Google Scholar, Microsoft Academic, Semantic Scholar) and third-party databases (Dimensions, The Lens, Scilit, OpenAlex). The first ones show clear problems with the number of fields that describe publications and the completeness of them. This could be due to these databases mainly obtain their data 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 their publications; it is the services with the lowest proportion of open access documents (35.4%) and external links (39.1%); almost 60% of publications do not have a document type; and it is the database with the most publications without publication date (57.9%). These figures evidence that Semantic Scholar has serious problems processing bibliographic information, which would cause the low quality of their metadata. In addition, the high proportion of records with Microsoft Academic's id (97.8%) suggests that the core of Semantic Scholar relies on Microsoft Academic, and not so much 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 the half of the publications include this information (Visser et al., 2021; Färber & Ao, 2022). This difference with our results could be due to patents are not included in our study (approx. 20%). However, it highlights in the coverage of external links (80%). These results are important to understand how other products based on their data (Semantic Scholar, The Lens and OpenAlex) have inherited or solved these problems.

The greatest problem of Google Scholar is that provides very little information about publications. Basic information such as document type, bibliographic information, publication date or identifiers is missing in this database. Excepting citations and versions, Google Scholar barely adds value to their records. However, as search engine is the service that provides the most external links (97.1%) and detects the most open versions (52.2%), which confirms that Google Scholar is the best gateway to access scientific literature on the Web (Martín-Martín et al., 2021).

In an opposite way, third-party databases supported on Crossref (Dimensions, The Lens and Scilit) have more metadata details and with a high completeness degree. Dimensions could be considered the database that enriches and improves the most the information provided by Crossref. It is the product that indexes the most publication's abstract (69.6%), identifies the most open access articles (44.5%) (Basson et al., 2022); the best coverage of bibliographic info (volume=100%, issue=81%, pages=92%), and 100% of publications with typology. These results illustrate that Dimensions makes a great effort to improve Crossref's metadata, adding abstracts, document typology, open access status, etc., to their records with a high completeness rate. However, other sources such as Scilit and The Lens shows 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%). The findings expose that the main risk of third-party databases based on external sources is that they require a great processing effort to improve the quality of their metadata. A similar case is OpenAlex, which is based both on Microsoft Academic and Crossref. This integration of different sources would cause losing of information, with a high proportion of missing bibliographic data (volume=62%, issue=50%, pages=51%). Results about OpenAlex allow us to suggest 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's 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 to offer a reliable scholarly bibliographic product (Priem, et al., 2022).

7. Conclusions

The obtained results allow us to conclude that the use of a random Crossref's sample makes possible the comparison of a wide range of scholarly bibliographic databases, benchmarking the information amount and completeness degree of these databases with regard to different facets. This method has hence measured the number of publications with abstract, external links, open access status, document type or publication date in comparison with the remaining databases. Even, the extraction of the same data in each database has favored the observation of overlaps, that have led us to identify possible connections between databases.

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

Finally, Dimensions is the product that 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 they display some integration problems with lack of information and low completeness rates in specific fields. Contrarily, search engines such as Semantic Scholar and Google Scholar lack of important fields for identifying and searching publications (document types, some bibliographic info). Microsoft Academic is the search engine that provide the most publication fields and their completeness rate is high, although it lacks of information on open access status and document type for some publications.

8. Competing interests statement

José Luis Ortega is member of the Scilit advisory board, which made possible accessing to API documentation.

9. 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.

10. Data availability statement

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 decided to upload 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.

11. 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.

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. doi: 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.

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*, 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

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

Herrmannova, D., & Knoth, P. (2016). An analysis of the microsoft academic graph. *D-lib Magazine*, *22*(9/10), 37.

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

Hug, S. E., & Brändle, M. P. (2017). The coverage of Microsoft Academic: Analyzing the publication output of a university. *Scientometrics*, *113*(3), 1551-1571.

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.

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.

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

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

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 preprint 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).

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.

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 [on line]. 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* (pp. 739-739).

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.