Comparing bibliographic descriptions in seven free-access databases

Lorena Delgado-Quirós* and José Luis Ortega*

*ldelgado@iesa.csic.es; jortega@iesa,csic,es
0000-0001-8738-7276; 0000-0001-9857-1511
Institute for Advanced Social Studies (IESA), Spanish National Research Council (CSIC), Spain

This communication aims to analyse the information that a large set of free-access databases (i.e., Crossref, Dimensions, Microsoft Academic, OpenAlex, Scilit, Semantic Scholar, The Lens) provides about indexed publications in their databases. Using a random sample of 116k publications from Crossref, each database was queried to retrieve the same document list with the purpose of comparing the metadata of their publications. The results show that the completeness degree is different between databases and that the search engines show more problems to extract abstracts and assign document typologies. Dimensions is the product that obtain the highest completeness percentages in abstracts, open access documents, bibliographic data and document types.

1. Introduction

The recent proliferation of bibliographic scholarly databases has stimulated their interest, mainly regarding to their possibilities to find scientific literature and provide different bibliometric indicators. This interest has been expressed in several studies that have tested the performance of these new systems in relation to traditional 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 calculating *ad hoc* bibliometric indicators. On the other hand, they are similar to search engines because they opt by a free access model in which users do not require subscription fee to search and retrieve documents.

However, these hybrid products have the particularities 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.

However, this multiple and varied availability of bibliographic data also presents a challenge because the integration of these data from different sources requires intense 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 metadata of 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 providing 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. Methods

2.1 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. Seven bibliographic databases were considered for the study:

Crossref, Dimensions, The 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.
- They could provide a reliable endpoint (e.g. Rest APIs, dump files) to extract information about their metadata.
- They also provide metrics for research evaluation.

2.2 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, favouring 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 selection 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.

2.3 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. 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) 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.
- **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.
- **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/) in January 2023 to supply this limitation.

- **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, 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 Microsoft Academic (https://aka.ms/msracad) 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 R package was used to extract the data. However, API was directly queried after to detect some problems in the retrieval process.

This study has a qualitative-quantitative approach, in which we extract large data samples from different sources to then compare the quality of the included information. Due to this we have analyzed the API documentation of each platforms to know which fields are available and what information contain each one.

3. 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 improve the identification and retrieval of these items, and their connection with other entities. Due to this, publication is the entity with more fields, going from the 38 fields in Crossref to the 18 in Semantic Scholar. Next, we analyse the fields used by each database to describe the main characteristics of a publication.

3.1. Abstract

This is an important access point to the content of the publication because it provides a summary of the research. All the analysed databases index this element. In the case of Microsoft Academic, the table with the abstract is not already available and this information could not be retrieved.

Table 1. Proportion of articles with abstract in each database

Databases	fields	Samples	Completeness	Completeness %
Crossref	abstract	116,592	15,927	13.66%
Dimensions	abstract	105,062	73,145	69.62%
The Lens	abstract	4,996	3,133	62.7%
Scilit	abstract	113,422	57,300	50.52%
Semantic Scholar	abstract	92,314	50,263	54.45%
OpenAlex	abstract_inverted_index	115,881	73,899	63.77%

Table 1 shows the proportion of publications with abstract in each database. Dimensions is the database that indexes more articles with abstract (69.6%), followed by OpenAlex (63.8%) with similar proportions. 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 that do not always include a formal abstract. This low percentage of abstracts in Crossref show 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%).

3.2. 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 four fields (*landing_page_url*, *pdf_url*, *host_venues_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 *host_venues_url* and *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

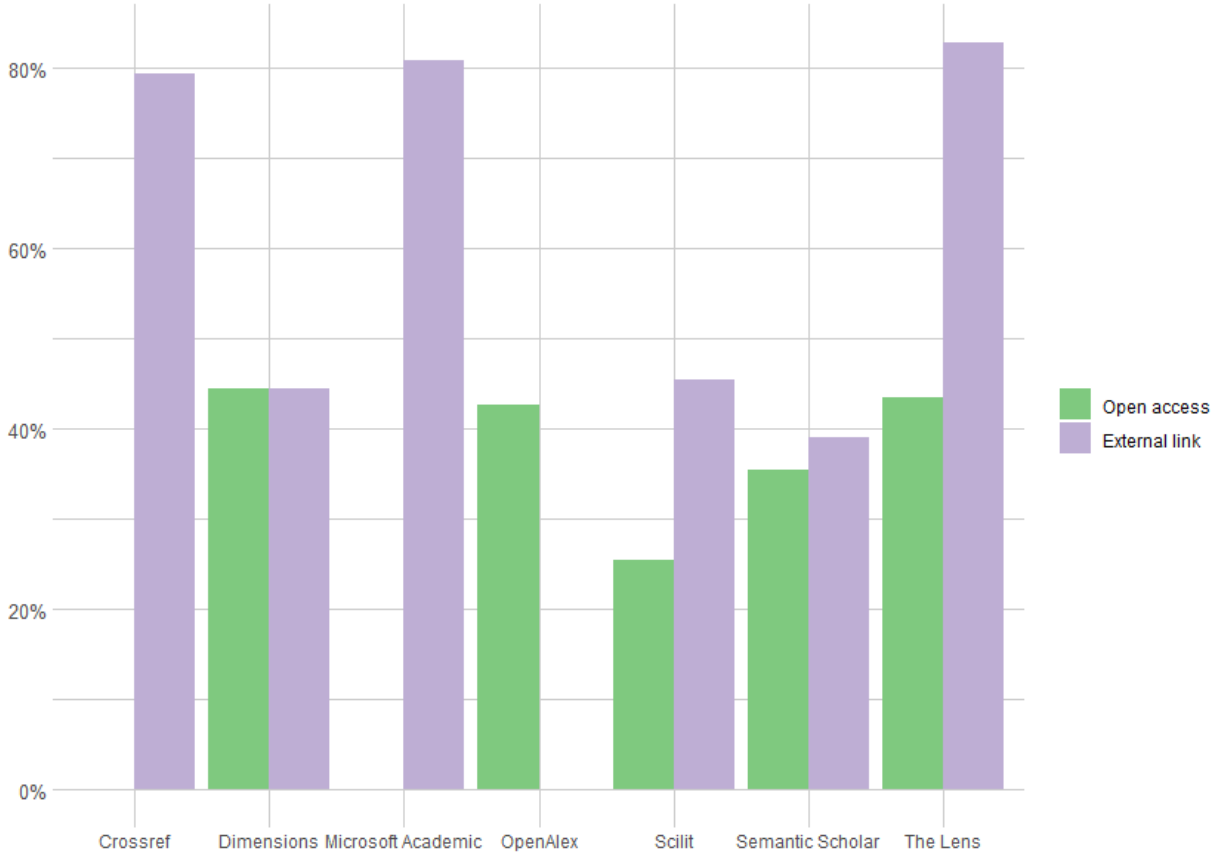

Figure 1 depicts the percentage of publications with external links to the original source and information about if they are open access or not. The Lens (82.9%), Microsoft Academic (80.8%) and Crossref (79%) are the databases that include more external links. The great coverage of external links by Microsoft Academic is due to, as academic search engine, crawl

the Web extracting urls from research publications. 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, 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, Dimensions (44.5%), The Lens (43.6%) and OpenAlex (42.6%) are the databases that identify more open access publications, while Semantic Scholar (35.4%) and Scilit (25.4%) capture fewer open documents.

3.3. Bibliographic information
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 do not have a field for issue.

Figure 2. Proportion of bibliographic records with information about volume, pages and issue

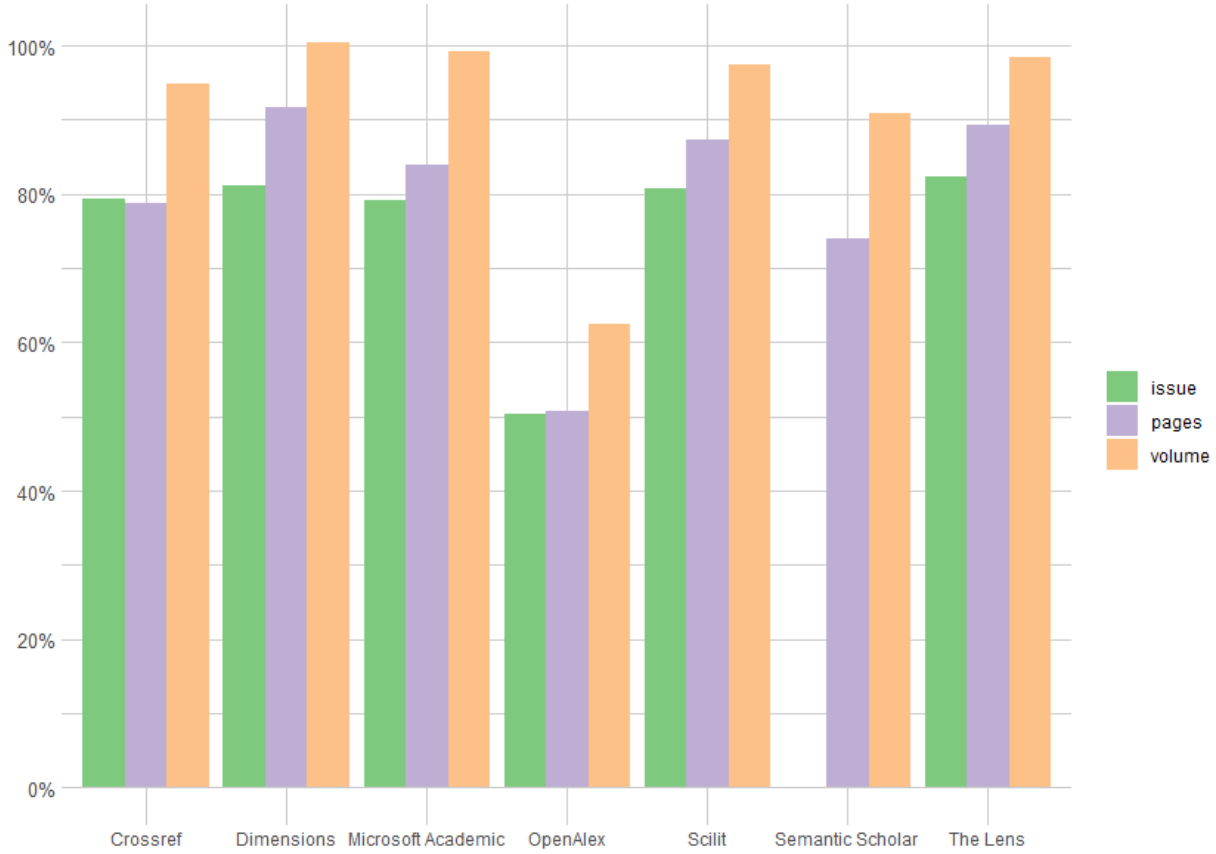

Figure 2 depicts the proportion of bibliographic data in each database for journal articles. In general, all the databases show high rates of completeness, including more information about volume than pages and issues. In this sense, Dimensions is again the platform that has highest completeness rates 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. Even more, if we assume that this database should be similar to Microsoft Academic. A manual inspection

confirmed this lack of data, in which almost all the records ingested in December 2022 do not include this information.

3.4. Document typology

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 2).

Table 2. Number of document typologies and completeness degree in each database

Source	**Typologies**	**Samples**	**Completeness**	**Completeness**
Crossref	33	116,592	116,592	100%
Dimensions	6	105,062	105,062	100%
Microsoft Academic	7	92,124	74,577	80.95%
The Lens	17	115,570	115,396	99.85%
Scilit	20	113,422	113,168	99.78%
Semantic Scholar	12	91,370	38,096	41.69%
OpenAlex	33	115,881	115,853	99.98%

Table 2 displays the number of different document types and the number of records categorized in each database. 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 (81%) 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 in content criteria (Review, Study, CaseReport, etc.) than in formal ones.

Figure 3. Alluvial graph with the transfer of document types between Crossref and the other databases

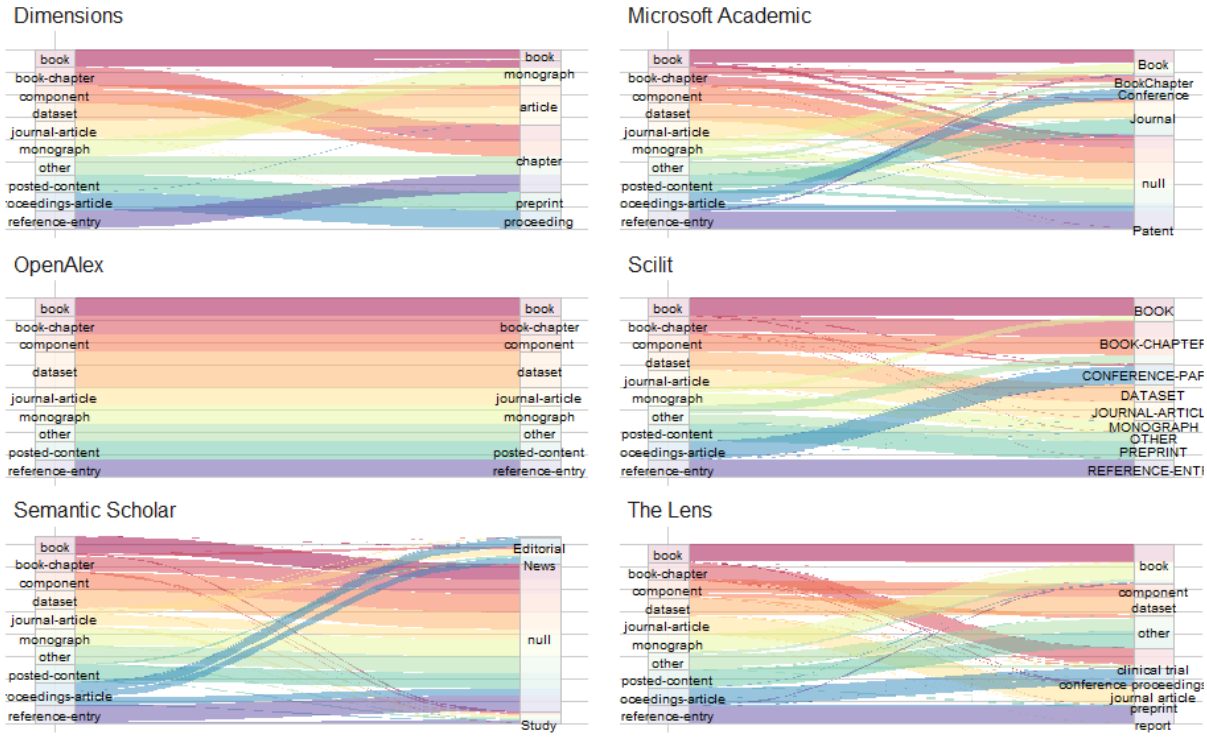

Figure 3 shows different alluvial graphs illustrating the document types transfers between Crossref classification and the systems of each database. The aim is to elucidate how each database assign document types to their records. To improve the clarity of the graph, only the ten most frequent categories in Crossref were displayed. For instance, Dimensions reduces significantly the document categories, integrating book-chapter (99.9%), component (78.7%), other (100%) and reference-entry (100%) in chapter category, and dataset (100%) and journal-article (99.1%) in article. Microsoft showed 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, assigning correctly 46.2% of proceeding articles and 35.3% of journal articles. 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%).

4. Conclusions

This descriptive and comparative analysis of the publication metadata supplied by the seven most relevant and accessible scholarly databases has reported important insights about how these services ingest and process the information that they publicly provide. Four parameters (abstract, external links, bibliographic information and documents typology) were comparatively analysed to describe the performance of each database. In general, we can conclude that Dimensions is the service that has the best metadata quality and completeness, because is the database that indexes most abstracts (69.6%), identifies most open access publications (44.5%) (Basson et al., 2022), has the best completeness of volume and pages, and 100% of records are assigned to a typology. Contrarily, Semantic Scholar presents important problems in the indexation of abstracts (54.5%), the assignation of document typologies (41.7%) and the inclusion of external links. We also can conclude that hybrid products show better metadata quality than academic search engines, especially with the classification of document types and indexation of abstracts. These differences call into

question the reliability of search engines to extract bibliographic information directly from websites and classify documents according to this information.

Open science practices

This communication is based on open (Crossref, OpenAlex, Microsoft Academic) and proprietary sources (Dimensions, Scilit, Semantic Scholar, The Lens). Due to this, raw data from proprietary sources cannot be publicly released.

Author contributions

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

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

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.

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.

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

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/