Philippe Mongeon
Dalhousie University, Halifax, NS, Canada

Madelaine Hare
University of Ottawa, Ottawa, ON, Canada

Geoff Krause
Dalhousie University, Halifax, NS, Canada

Rebecca Marjoram
Dalhousie University, Halifax, NS, Canada

Poppy Riddle
Dalhousie University, Halifax, NS, Canada

Rémi Toupin
Dalhousie University, Halifax, NS, Canada

Summer Wilson
Dalhousie University, Halifax, NS, Canada

INVESTIGATING DOCUMENT TYPE DISCREPANCIES BETWEEN OPENALEX AND THE WEB OF SCIENCE

Abstract

Bibliometrics, whether used for research or research evaluation, relies on large multidisciplinary databases of research outputs and citation indices. The Web of Science (WoS) was the main supporting infrastructure of the field for more than 30 years until several new competitors emerged. OpenAlex, launched in 2022, stands out for its openness and extensive coverage. While OpenAlex may reduce or eliminate barriers to accessing bibliometric data, one of the concerns that hinder its broader adoption for research and research evaluation is the quality of its metadata. This study aims to assess the metadata quality of works in OpenAlex and WoS, focusing on document type accuracy. We observe that over 4% of the publications indexed in both OpenAlex and WoS appear to be misclassified as research articles or reviews, and that the vast majority (about 97%) of these errors occur in OpenAlex. By addressing discrepancies and misattributions in document types this research seeks to enhance awareness of data quality issues that could impact bibliometric research and evaluation outcomes.

Keywords

Bibliometrics; research evaluation; OpenAlex; Web of Science; Library and Information Science; open data; metadata; document type

Introduction

Bibliometrics has been used in research and research evaluation for about half a century (Narin, 1976), supported by Eugene Garfield's development of what we now know as the Web of Science (WoS). It took more than 30 years for competitors and other players to emerge: Elsevier's Scopus was founded in 1996, Crossref in 1999, Google Scholar in 2004, Microsoft Academic and Dimensions in 2018, and OpenAlex in 2022.

Bibliometrics has long paid attention to the coverage and data quality of these data sources and investigated differences in evaluative outcomes produced from them. The advent of OpenAlex, an open bibliographic database that indexes over 250 million scholarly works with broader coverage of the Humanities, non-English languages, and the Global South than traditional indexes (Priem et al., 2022), generated a new wave of such studies. As an aggregator of data from Microsoft Academic Graph (MAG), Crossref, ORCID, ROR, DOAJ, Unpaywall, Pubmed, and multiple other sources, OpenAlex's data coverage and quality have been of interest in utilizing it for quantitative work. This scrutiny has resulted in observations that using the subset of OpenAlex works indexed in WoS or Scopus might contain incomplete, or erroneous metadata. The Leiden Ranking Open Edition published by the Centre for Science and Technology Studies (CWTS) provides credence to the claim as it uses only a subset of OpenAlex data that is selected based on criteria similar to those used for inclusion in the WoS[1].

The coverage of OpenAlex in relation to established bibliographic databases has received attention since its emergence. Culbert et al. (2024) investigated the coverage of reference items between OpenAlex, WoS, and Scopus. They found that OpenAlex was comparable with commercial databases from an internal reference coverage perspective if restricted to a core corpus of publications similar to the other two sources, though it lacked cited references. Low reference, funder, and affiliation coverage was also found by Alonso-Alvarez and van Eck (2024), though they observed OpenAlex's coverage of publication and author information was high compared to WoS and Scopus. Simard et al. (2024) and Maddi et al. (2024) investigated the OA journal coverage of OpenAlex, WoS, and Scopus using the DOAJ and ROAD databases as reference databases. Both studies found that OpenAlex indexes more journals and provides more balanced geographical coverage. Céspedes et al. (2024) determined that OpenAlex's linguistic coverage (75% English) far surpassed that of WoS (95% English) from the metadata, with the former reduced to 68% upon manual verification of the works themselves. However, the language field in OpenAlex is algorithmically detected from the title and abstract metadata, introducing limitations[2]. In a coverage comparison of six databases, Ortega and Delgado-Quirós (2024) found OpenAlex indexes more retracted works than WoS, Scopus, and PubMed.

Other studies focused on the quality of OpenAlex metadata and showed that institutional metadata is missing from many OpenAlex records (Bordignon, 2024; Zhang et al., 2024) and funding metadata is also lacking (Schares, 2024). Haupka et al. (2024) observed a broader range of materials as research publications in OpenAlex compared to Scopus, WoS and PubMed,

[1] https://open.leidenranking.com/information/indicators

[2] https://docs.openalex.org/api-entities/works/work-object#language

potentially explained by Ortega and Delgado-Quirós (2024) as resulting from the database's reliance on Crossref's less precise system of classification.

While OpenAlex's coverage has been under intense investigation, more needs to be understood about the comparableness of its metadata to other data sources. Our work thus aims to assess the metadata quality of works in both OpenAlex and WoS. This work in progress focuses on one specific metadata element: the document type. More specifically, it addresses the following research questions (RQs):

> **RQ1.** How are WoS and OpenAlexrecords distributed across document types?
>
> **RQ2.** What is the share of OpenAlex records with a document type discrepancy with the matching WoS record?
>
> **RQ3.** How frequent is the misattribution of the article or review document type to records in WoS or OpenAlex?

Since most usage of the WoS or OpenAlex databases in bibliometric research and evaluation is typically limited to research articles and reviews, investigating the accuracy of document types in traditional and emerging databases is an important step to raise awareness of data quality issues that could affect findings.

Methods

Data collection

The WoS data used in this study was retrieved from a relational database version of the WoS hosted by the Observatoire des sciences et des technologies (OST) and limited to the Science Citation Index (SCI), the Social Sciences Citation Index (SSCI), and the Arts & Humanities Citation Index (A&HCI). In the OST database, every journal is assigned to one of 143 specialties of the NSF classification.

We collected all WoS records with a DOI published between 2021 and 2023 (N = 7,645,000). We removed 16,373 (0.2%) WoS records with multiple document types to avoid complications with the analysis. Of the remaining 7,628,627 WoS records, 6,594,747 (85.2%) had a DOI match in the February 2024 snapshot of OpenAlex accessed through Google Big Query (see Mazoni & Costas, 2024). We used these 6,594,747 records for our analysis.

Data analysis

We then compared the document type indicated in WoS and OpenAlex to identify discrepancies. We are mainly interested in identifying erroneous inclusions of records in bibliometric analyses typically limited to articles and reviews. Therefore, we do not consider discrepancies where a record is a review according to WoS and an article according to OpenAlex. Furthermore, OpenAlex indexes conference papers as articles and the source type (conference) is meant to distinguish them from journal articles. For these reasons, we only analyzed discrepancies for which the record is identified as an article or a review in either WoS or OpenAlex and identified as neither an article nor a review in the other source. We also excluded discrepancies in which the record is identified as a meeting abstract in WoS and an article in OpenAlex. Overall, we found 311,220 discrepancies that met these criteria, which amounts to 4.6% of all records in the dataset.

Results

Table 1 and Table 2 present the distribution of records across document types in WoS and OpenAlex, respectively, to provide a general picture of the databases' content and of the differences in their classification. While WoS contains twice as many document types as OpenAlex, these differences appear mainly among the less frequent types, in line with the findings of Haupka et al. (2024). The vast majority of documents in both data sources are articles and reviews.

Table 1. Number and share of records by document type in the Web of Science.

Document type	Number of records	% of records
article	5,424,938	82.2
review	509,515	7.7
editorial material	265,878	4.0
meeting abstract	117,134	1.8
letter	113,418	1.7
book review	67,304	1.0
correction	67,167	1.0
news item	11,974	0.2
retraction	8,277	0.1
biographical-item	6,481	0.1
cc meeting heading	850	< 0.05
poetry	776	< 0.05
expression of concern	623	< 0.05
reprint	379	< 0.05
art exhibit review	250	< 0.05
item withdrawal	209	< 0.05
film review	174	< 0.05
bibliography	124	< 0.05
fiction, creative prose	108	< 0.05
theater review	92	< 0.05
record review	35	< 0.05
music performance review	11	< 0.05
software review	11	< 0.05
music score review	7	< 0.05
hardware review	5	< 0.05
tv review, radio review*	5	< 0.05
dance performance review	4	< 0.05
excerpt	3	< 0.05

Document type	Number of records	% of records
database review	2	< 0.05
data paper	1	< 0.05
Note*	1	< 0.05
script	1	< 0.05

**Note* and *TV Review*, *Radio Review*, *Video Review* were retired as document types and are no longer added to items indexed in the WoS Core Collection. They are still usable for searching or refining/analyzing search results.[3]

Table 2. Number and share of records by document type in OpenAlex.

Document type	Number of records	% of records
article	5,832,410	88.4
review	511,706	7.8
letter	136,056	2.1
editorial	59,649	0.9
erratum	48,178	0.7
retraction	4,706	0.1
book-chapter	1,333	< 0.05
preprint	1,223	< 0.05
paratext	303	< 0.05
book	95	< 0.05
report	69	< 0.05
dataset	13	< 0.05
other	9	< 0.05
dissertation	4	< 0.05
supplementary-materials	2	< 0.05
reference-entry	1	< 0.05

Discrepancies in document types

Tables 3 and 4 show that the vast majority (N = 301,884, 97%) of the 311,220 discrepancies found are cases where a record is an article or review in OpenAlex but not WoS. Based on the verified sample, approximately 261,733 (86.7%) would be erroneous in OpenAlex. On the other hand, we found only 9,336 cases where the record is an article or review in WoS but not OpenAlex. Based on the verified sample, we estimate that about 5,406 (57.9%) of these records would be erroneous in WoS.

Table 3. Frequency distribution of discrepancies for articles and reviews in WoS.

[3] https://webofscience.help.clarivate.com/en-us/Content/document-types.html

Document type in OpenAlex	**Number of discrepancies**	**% of discrepancies**	**% true errors based on sample**
letter	5,004	53.6	45.8
editorial	1,341	14.4	90.9
book-chapter	1,303	14.0	100.0
preprint	1,182	12.7	0.0
paratext	239	2.6	100.0
erratum	149	1.6	n/a
report	68	0.7	n/a
retraction	28	0.3	n/a
dataset	13	0.1	n/a
book	5	0.1	n/a
other	2	<0.05	n/a
dissertation	1	<0.05	n/a
supplementary-materials	1	<0.05	n/a
Overall	9,336	100	57.9

Table 4. Frequency distribution of discrepancies for articles and reviews in OpenAlex.

Document type in WoS	**Number of discrepancies**	**% of discrepancies**	**% true errors based on sample**
editorial material	160,807	53.3	78.1
book review	67,169	15.2	98.0
letter	30,242	8.0	93.6
correction	19,208	5.6	98.4
news item	11,327	3.5	95.5
biographical-item	5,906	1.9	98.2
retraction	3,712	1.2	100.0
cc meeting heading	842	0.3	100.0
poetry	750	0.2	83.3
expression of concern	574	0.2	100.0
reprint	372	0.1	100.0
art exhibit review	250	0.1	50.0
film review	174	0.1	100.0
item withdrawal	145	0.0	n/a
bibliography	120	0.0	n/a
fiction, creative prose	108	0.0	n/a
theater review	92	0.0	n/a
record review	35	0.0	n/a
music performance review	11	0.0	n/a
software review	11	0.0	n/a
music score review	7	0.0	n/a
hardware review	5	0.0	n/a

Document type in WoS	Number of discrepancies	% of discrepancies	% true errors based on sample
tv review, radio review	5	0.0	n/a
dance performance review	4	0.0	n/a
excerpt	3	0.0	n/a
database review	2	0.0	n/a
data paper	1	0.0	n/a
note	1	0.0	n/a
script	1	0.0	n/a
Overall	301,884	100	86.7

Discussion and conclusion

The value of bibliographic data sources is derived from different elements, including their coverage, completeness, and data accuracy (Visser et al., 2021). This research contributes to understanding the value and utility of OpenAlex as a data source by investigating its metadata discrepancies in relation to document type. Our findings support those of past studies that found metadata quality in OpenAlex to need improvement compared to WoS. Accurate document type classifications are critical for bibliometric research and evaluation, and calibrating diverse document types across disciplines and databases remains a challenge (Haupka et al., 2024). For the next stage of this research, we will include additional metadata elements widely used in bibliometric analyses and investigate disciplinary differences in metadata quality. Further research will examine how metadata quality issues in OpenAlex could affect journal and institutional-level metrics and, thus, the results of institutional rankings like the open edition of the Leiden Ranking. The Paris Conference on Open Research Information and the Barcelona Declaration on Open Research Information are two initiatives that impressed the need for and normalization of open research information. The Barcelona Declaration called for signatories to work with systems that support open research information (Barcelona Declaration, 2024). With the tide turning toward open data sources and researchers and institutions embracing OpenAlex and other open data sources and tools, more research will be needed on the quality and coverage of OpenAlex and the other data sources it depends on.

References

Alonso-Alvarez, P., & Eck, N. J. van. (2024). *Coverage and metadata availability of African publications in OpenAlex: A comparative analysis* (No. arXiv:2409.01120). arXiv. https://doi.org/10.48550/arXiv.2409.01120

Alperin, J. P., Portenoy, J., Demes, K., Larivière, V., & Haustein, S. (2024). *An analysis of the suitability of OpenAlex for bibliometric analyses*. arXiv. https://doi.org/10.48550/arXiv.2404.17663

Barcelona Declaration. (2024). “Barcelona Declaration on Open Research Information”. https://barcelona-declaration.org/

Bordignon, F. (2024). *Is OpenAlex a revolution or a challenge for bibliometrics/bibliometricians?* https://enpc.hal.science/hal-04520837

Céspedes, L., Kozlowski, D., Pradier, C., Sainte-Marie, M. H., Shokida, N. S., Benz, P., Poitras, C., Ninkov, A. B., Ebrahimy, S., Ayeni, P., Filali, S., Li, B., & Larivière, V. (2024). *Evaluating the Linguistic Coverage of OpenAlex: An Assessment of Metadata Accuracy and Completeness*. arXiv. https://doi.org/10.48550/arXiv.2409.10633

Culbert, J. H., Hobert, A., Jahn, N., Haupka, N., Schmidt, M., Donner, P., & Mayr, P. (2024). *Reference Coverage Analysis of OpenAlex compared to Web of Science and Scopus*. Cornell University. https://doi.org/10.48550/arxiv.2401.16359

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

Garfield, E., & Sher, I. H. (1963). New factors in the evaluation of scientific literature through citation indexing. *American Documentation*, *14*(3), 195–201. https://doi.org/10.1002/asi.5090140304

Haupka, N., Culbert, J.H., Schniedermann, A., Jahn, N., & Mayr, P. (2024). Analysis of the publication and document types in OpenAlex, Web of Science, Scopus, Pubmed and Semantic Scholar. *ArXiv*. https://doi.org/10.48550/arXiv.2406.15154

Maddi, A., Maisonobe, M., & Boukacem-Zeghmouri, C. (2024). Geographical and disciplinary coverage of open access journals: OpenAlex, Scopus and WoS. *ArXiv*. https://doi.org/10.48550/arXiv.2411.03325

Mazoni, A., & Costas, R. (2024). *Towards the democratisation of open research information for scientometrics and science policy: the Campinas experience*. https://www.leidenmadtrics.nl/articles/towards-the-democratisation-of-open-research-information-for-scientometrics-and-science-policy-the-campinas-experience

Narin, F. (1976). *Evaluative bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity*. Computer Horizons Washington, D. C.

Ortega, J. L., & Delgado-Quirós, L. (2024). The indexation of retracted literature in seven principal scholarly databases: A coverage comparison of dimensions, OpenAlex, PubMed, Scilit, Scopus, The Lens and Web of Science. *Scientometrics*, *129*(7), 3769–3785. https://doi.org/10.1007/s11192-024-05034-y

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

Schares, E. (2024). *Comparing Funder Metadata in OpenAlex and Dimensions*. https://doi.org/10.31274/b8136f97.ccc3dae4

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

Shi, J., Nason, M., Tullney, M., & Alperin, J. (2025). Identifying metadata quality issues across cultures. *College & Research Libraries*, *86*(1). https://doi.org/10.5860/crl.86.1.101

Simard, M.-A., Basson, I., Hare, M., Lariviere, V., & Mongeon, P. (2024). *The open access coverage of OpenAlex, Scopus and Web of Science*. arXiv. https://doi.org/10.48550/arXiv.2404.01985

van Eck, N. J., Waltman, L., & Neijssel, M. (2024, October 9). Launch of the CWTS Leiden Ranking Open Edition 2024. *Leiden Madtrics.* https://www.leidenmadtrics.nl/articles/launch-of-the-cwts-leiden-ranking-open-edition-2024

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

Zhang, L., Cao, Z., Shang, Y., Sivertsen, G., & Huang, Y. (2024). Missing institutions in OpenAlex: Possible reasons, implications, and solutions. *Scientometrics*. https://doi.org/10.1007/s11192-023-04923-y